DOCUMENT RESUME

ED 206 842 CE 029 927

AUTHOR Tallmadge, G. Kasten; Yuen, Sandra D.
TITLE Study of the Career Intern Program. Final Report--Task B: Assessment of Intern Outcomes.
INSTITUTION RMC Research Corp., Mountain View, Calif.
SPONS AGENCY National Inst. of Education (ED), Washington, D.C.
REPORT NO RMC-UR-482
PUB DATE May 81
CONTRACT 400-78-0021
NOTE 180p.; For related documents see CE 029 925-930 and TM 810 654.

EDRS PRICE MF01/PC08 Plus Postage.
DESCRIPTORS *Achievement Tests; *Career Education; Case Studies; Dropout Prevention; Dropout Programs; *Dropouts; *Economically Disadvantaged; Experiential Learning; *Field Experience Programs; High School Equivalency Programs; High School Students; Nontraditional Education; *Outcomes of Education; Potential Dropouts; Program Descriptions; *Program Effectiveness; Secondary Education; Student Attrition; Student Characteristics; Student Educational Objectives; Student Evaluation; Success
IDENTIFIERS *Career Intern Program

ABSTRACT

A study assessed the impact of the Career Intern Program (CIP) on participating students. (The CIP is an alternative high school designed to enable disadvantaged and alienated dropouts or potential dropouts to earn regular high school diplomas, to prepare them for meaningful employment or postsecondary education, and to facilitate their transition from school to work by providing instruction, counseling, hands on career exposure, diagnosis/assessment, and climate.) To evaluate student outcomes, standardized reading and mathematics achievement tests were administered to both an experimental and a control group on four occasions (upon entering the program, six and twelve months thereafter, and six to twelve months after completing the program). The declining number of students in the test samples (1680 students tested initially, 786 students tested midway into the program, and 500 tested at its conclusion) reflected the program's high attrition rates. Despite the high attrition rate (which may be explained, at least in part, by a number of operational problems involving tight scheduling, funding, and unrealistic enrollment quotas), achievement test results support the success of CIP. (Related reports evaluating other aspects of CIP are available separately through ERIC--see note.) (HN)

RMC Report No. UR 482

STUDY OF THE CAREER INTERN PROGRAM

Final Report Task B:

Assessment of Intern Outcomes

G. Kasten Tallmadge
Sandra D. Yuen

May 1981

Prepared for the
National Institute of Education

RMC Research Corporation
Mountain View, California

The research reported herein was performed pursuant to Contract No. 400-78-0021 with the National Institute of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Institute position or policy, and no official endorsement by the sponsor should be inferred.

Contents

List of Tables

Page

Page

List of Figures

PREFACE

This report is concerned with the impact that the Career Intern Program has had on participating students. It is a traditional outcome evaluation, heavily quantitative in its orientation. Unfortunately, it serves well to illustrate the limitations that traditional experimental approaches have when applied to social reform programs in field settings. The various designs that were employed had to be adapted to the practicalities of real-world conditions, experimental controls were inadequate, and attrition from all groups studied was high. In the end, important assumptions underlying statistical tests were badly violated and serious concerns arose as to the internal validity of all of the analyses that were undertaken.

In putting the report together, we have tried to point out the many flaws that exist. At the same time we have attempted to salvage what is useful and to piece together the various bits of evidence that have been assembled in as meaningful a way as possible. In doing so, we have tried to tie observed outcomes to significant implementation events that took place at each of the four program sites. Some of the inferences we have drawn are quite speculative, others are more defensible. Throughout our efforts, however, we were frustrated by the inadequacy of the tools we had to use.

Our frustration was not unexpected. We had seen the evaluation of the CIP prototype and were aware that we would encounter even greater problems. We were also aware, as Cronbach, Ambron, Dornbusch, Hess, Hornik, Phillips, Walker, and Weiner (1980) have noted, that "Few evaluative experiments to date have achieved all the following earmarks of internal validity: genuinely randomized assignment; meaningful, describable treatments; samples large enough to give reasonable statistical power; and attrition low enough to maintain the initial equivalence" (p. 308).

The fact that the problems we encountered in this study were not unique failed to make us feel much better because the report does not adequately reflect what we believe we know about the program. For approximately three years, members of the RMC project staff have spent considerable time on site, have had lengthy conversations with staff and students, and have observed all aspects of program operations. Based on these experiences we believe that the CIP, when properly implemented, is a powerful force for reshaping the lives of disadvantaged and alienated youths. We believe that program participants realize cognitive achievement benefits and develop useful career awareness. We believe that more of them graduate from high school, go on to further education, and/or obtain meaningful employment than would be the case without the CIP. The evidence contained in this report, however, while supportive of these beliefs, is not entirely conclusive.

In 1979, Donald Campbell took the position that, where qualitative data collected through interviews and observations "are contrary to the quantitative results, the quantitative results should be regarded as suspect" (p. 53). In the case of the present study, the quantitative and qualitative data are in general agreement. The problem lies solely in the fact that the quantitative data are vulnerable to attacks regarding their internal validity. Following Campbell's lead, we now take the position that the credibility of the quantitative findings is substantially enhanced by the fact that the qualitative data also support program success.

Presentation of all the qualitative data that support program success is beyond the scope of this report. It is thoroughly documented in a companion volume (Fetterman, 1981), however, to which the interested reader is referred.

In this report, we have advanced several hypotheses that may appear to be inadequately supported by the available data. In most instances, the cited Fetterman report contains additional relevant information. Even so, some of our inferences may go beyond the data. We were guided by the following statement:

> Social scientists are trained to suppress relationships that do not reach statistical significance. However, no relation that makes sense ought to be discarded. We say this despite the truism that an explanation can be dreamed up to fit any adventitious result. (Cronbach et al., 1980, p. 315).

We hope and believe that we have not "dreamed up" explanations to fit the data. At the same time, we are aware that the "hard" data do not, in and of themselves, provide conclusive proof that the Career Intern Program was successful in achieving its objectives. It is only when one considers the qualitative data as well that the argument seems to us to be overwhelmingly convincing.

ACKNOWLEDGEMENTS

The authors are indebted to many individuals for help related to this report. German Calder, Susie Guiora, Thomas Hyde, Patrick Lennahan, Roberta Staples, Gail Sydnor, and Robert Vaughan bore primary responsibility for on-site data collection. Linda Terhune, Fred Weiner, Mary Pat Gaspich, Susan Gaspich, and Kathleen Gaspich scored tests, coded data, and assisted with the statistical analyses. We are very grateful for all of these contributions.

The CIP directors and our colleagues at Opportunities Industrialization Centers of America, particularly Robert Jackson, also deserve many thanks for their cooperation, assistance, and understanding in conducting this evaluation. We are also grateful to Charles Stalford, Howard Lesnick, and Daniel Antonoplos of the National Institute of Education for their concern and guidance and for the encouragement they provided.

Finally, we wish to acknowledge the very helpful comments and suggestions of Robert Boruch and Andrew Porter who reviewed an earlier Task B report. This document is much improved as a result of the comments and suggestions they provided.

GKT
SDY

EXECUTIVE SUMMARY

Background of the Career Intern Program

The Career Intern Program (CIP) is an alternative high school designed to serve disadvantaged and alienated students (called interns) who either dropped out of regular high schools or who were considered potential dropouts. The objectives of the program are to enable students to earn a regular high school diploma (as opposed to a GED), to prepare them for meaningful employment, and to facilitate their transition from school to work. The program offers extensive counseling--academic, personal, and career--and attempts to make academic subjects palatable and relevant to the lives of the students through a heavy infusion of career-oriented content.

Run by a community-based organization, the Career Intern Program enjoys an unusual symbiotic working relationship with the local school district. It serves those students whose needs are not adequately met by the local high school, but the students remain on the local school's books. State monies that are distributed to the schools based on enrollment or attendance thus continue to flow to the local high school even though the students are being served by the CIP. The high schools award diplomas to students graduated by the CIP.

The CIP was initially developed in Philadelphia in the mid-1970s. An independent evaluation conducted by Richard A. Gibboney Associates (Gibboney Associates, 1977) found the program to be successful. The evidence of success was judged sound by the Joint (U.S. Office of Education and National Institute of Education) Dissemination Review Panel, and the program was approved by that group as eligible for federally funded dissemination.

Under authorization of the Youth Employment and Demonstration Projects Act (YEDPA, Public Law 95-93), the Department of Labor (DOL) and the National Institute of Education (NIE) entered into an Interagency Agreement in late 1977 to test the replicability of the CIP and to determine whether the same beneficial outcomes could be obtained in the replication sites. Subsequently, NIE contracted with the Opportunities Industrialization Centers of America (OIC/A) to manage the replication effort. OIC/A then, through a competitive bidding process, selected four local OIC chapters to undertake the CIP replication. Three of the selected sites were urban and one was located in a small (30,000) city.

Overview of the Evaluation

The work statement for the evaluation was prepared jointly by NIE and DOL. Four separate tasks were called for:

- Task A. Conduct studies and analyses as required to answer the questions, "What happens to the Career Intern Program in the process of implementation in additional sites? What factors account for the changes or adaptations, if any? For the fidelity, if any, to the original program goals and practices?" (RFP NIE-R-78-0004, p. 9)

- Task B. Conduct studies and analyses as required to answer the question, "Does the Career Intern Program continue to be effective in helping youth when it is implemented in sites other than the Philadelphia prototype?" (ibid, p. 13)

- Task C. Conduct studies and analyses as required to answer the question, "What happens to young people in the CIP program that could account for its effectiveness?" (ibid, p. 16)

- Task D. Conduct studies and analyses as required to answer the fourth question, "How does the CIP approach compare in effectiveness, feasibility, impact, and factors important for policy with other approaches undergoing comparable evaluations, to helping the population to be served through the Youth Employment Act?" (ibid, p. 20)

To assure comparability with the original CIP evaluation, the work statement specified that the evaluations of the replication sites employ the same instruments and designs as that study. While some modifications were eventually made to strengthen the study, care was taken to preserve the desired comparability.

The present report deals only with Task B. Task A and Task C, however, are highly relevant to the material presented herein as variations in the extent or manner in which individual components of the treatment were implemented almost certainly affected program outcomes. While an attempt has been made throughout this report to relate observed outcomes to implementation events and conditions, much more detailed information is provided in the reports of the other two tasks (Treadway, Stromquist, Fetterman, Foat, & Tallmadge, 1981; Fetterman, 1981).

Methodology

Social science research in the field cannot be implemented in strict accordance with the "rules" that govern laboratory studies. The primary problem for the present evaluation was very high attrition rates in both treatment and control groups. These high attrition rates rendered it impossible to determine with complete certainty whether observed differences between groups at posttest time resulted from the treatment or from some other influence (including attrition itself). This and other problems led the investigators to employ a variety of different evaluation approaches and data analyses strategies. By examining the data from several different

perspectives, it was reasoned, a more credible case could be made for the success or failure of the program in achieving its goals.

It is beyond the scope of this summary to describe each of the various techniques that was employed. Such descriptions are, of course, contained in the main body of the report. It should be noted here, however, that the different approaches yielded somewhat different results. Furthermore, since some of the assumptions underlying each approach were violated, it is not clear which "answer" (if any) should be believed. Lest too negative a picture be presented, however, we hasten to point out that the differences among results were not extreme and all tended to support the success of the CIP.

Implementation Events

When the CIP is well implemented, there is reason to expect that it will impact positively on participating students. When it is not well implemented, less sanguine expectations seem appropriate. It is important to make this point because each of the four CIP demonstrations experienced serious implementation difficulties at various times. Only meager evidence of success could reasonably be expected during these times.

One of the sites got off to a good start but then encountered serious difficulties that were never adequately resolved during the entire demonstration period. Another site that ran for many months in a truly exemplary manner fell into disarray when its director departed. Two other sites experienced severe start-up problems. One was well on the way to recovery when its director and several other key staff left. The other did achieve a high degree of implementation success--but not until the end of the demonstration period was imminent. Not one of the three cohorts of students studied at any of the four sites experienced a full year of program "treatment" unmarred by some sort of major trauma.

The fact that regularly attending students were often not receiving a "full" treatment was compounded by the irregular attendance of many others. In addition, some students were so poorly prepared academically that they simply could not cope with the curriculum and should never have been admitted to the program. Both of these problems were direct outgrowths of the extreme pressures applied to the sites to meet enrollment quotas.

Taken together, the various influences described above acted in a manner that could only detract from the measured impact of the CIPs. Still, when the programs were operating well there was ample evidence of success. Even when all was not well, some gains continued to be observed.

Results

Evidence was found that the CIPs had significant "holding power" over participating students. This holding power, furthermore, varied in direct proportion to the quality of program implementation. When all program components were in place and functioning smoothly, attendance was high and attrition was low. When the programs encountered implementation problems, attendance fell off and attrition increased.

In the area of reading achievement, results over the 12-month period between pre- and posttests, showed statistically significant gains when data were pooled across sites and cohorts. When the performance of CIP students was compared against expectations derived from normative data, however, the gain estimate was more than two-and-a-half times as large as that derived from the treatment-control comparison. While it is believed that the larger estimate is the more accurate one, some would argue that the smaller estimate was more credible. When the performance of CIP students was compared against the performance of students in other alternative programs, statistically and educationally significant advantages were found for the CIP.

Most of the invididual-site and individual-cohort gain estimates were statistically significant in the norm-referenced analyses. In the treatment-control analyses, only the across-site, across-cohort estimate attained significance.

In math, the picture was similar, but the gains were somewhat smaller. This finding was not surprising as all of the sites experienced great difficulty in attracting and retaining qualified math instructors. None of the pre-to-posttest gain estimates derived from treatment-control comparisons was statistically significant when "normal" analytic procedures (analyses of covariance) were used. Under an alternative approach (standardized gain analyses), a somewhat more positive picture emerged. In the norm-referenced analyses, statistically significant gains were found for all three of the cohorts studied when the data were pooled across sites.

Of the 12 individual-site, individual-cohort analyses, 5 showed statistically significant norm-referenced gains. Perhaps the most notable result of the math analyses was the fact that the gains were consistently positive at times when individual sites were known to have had appropriately qualified math teachers and consistently negative when they did not.

When the performance of CIP students was compared with that of students in other alternative high schools, the results strongly favored the CIP group at two individual sites and in the across-site analysis. The same results were obtained when CIP students were compared against a group of regular high school students.

Statistically significant gains were found on all three scales of the Career Development Inventory (Planning, Use of Resources, and Information) in several of the individual-site analyses. Across sites, the gain estimates were significant in over half of the cases. Gains on the Information scale, although statistically significant, were small. This finding was surprising in view of the heavy infusion of career-related material in the CIP curriculum. Examination of the scale's content, however, revealed that it was pitched at a global and theoretical level while the CIP's instruction was at a more job-specific, practical level.

Statistically significant gains in self-esteem were observed in half of the across-site analyses at posttest time. Interestingly, however, none of the corresponding analyses showed a significant treatment effect at midtest time. Other variables also showed smaller pre-to-midtest than pre-to-posttest effects, but in no other case was the difference so pronounced. It was concluded that changes in self-concept require extended exposure to the type of counseling and other program features offered by the CIP.

While it seemed logical to expect that CIP participants would experience an increased sense of control over their lives, scores on the Internal-External (locus of control) scale reflected significant gains in only a few scattered instances. Again, this inconsistency between impressions gained through extended on-site observations and the quantitative data was attributed to deficiencies in the instrument rather than failure of the treatment.

CIP participants and members of the control groups were followed up in the summer of 1980 and again in January and February of 1981. Analyses of the data obtained from these follow-ups are more directly related to the CIP's stated goals of helping participants earn their high school diplomas and enhancing their employability than those involving test scores. Gains on achievement, information, and self-concept tests may well be important, but they are at best intermediate goals of the program.

Comparisons between treatment and control groups in terms of the numbers who had graduated from high school, were currently enrolled, or had earned a GED were generally favorable.

For the fourth cohort, the high school status of the treatment group was significantly better than that of the control group at one individual site and across all four sites. This was despite the fact that serious implementation problems existed at one site. At that site, the status of the control group was better than that of the treatment group (although not significantly so).

The third-cohort data also showed a significant advantage for the treatment group over the control group at one individual site. The negative results of the site experiencing implementation

difficulties, however, prevented the differences from being significant across all four sites. When data were combined across the three sites that were not having implementation problems, a significant advantage was again found for the treatment group.

The second cohort had no control group. Across sites, however, a larger percentage of treatment group members had graduated from high school, were currently enrolled, or had earned a GED, however, than was the case with either the third or fourth cohorts. This relationship held at both the first and second follow-ups largely because the operational problems at one site that are referred to above had not yet developed.

The second stated goal of the CIP to which follow-up data were relevant was that of smoothing the transition from school to work. Because large numbers of students were still enrolled in school, however, it seemed most appropriate to compare treatment and control groups in terms of the numbers either in school or employed versus those not in school and not employed.

The results of these comparisons were slightly less favorable than those related to high school status, but still positive. The fourth-cohort treatment group presented a better picture than the control group at one individual site and across sites on the only follow-up that was conducted on that cohort. There were no significant differences between treatment and control groups for the third cohort, but members of the second and third cohorts who had participated in the program for at least three months were significantly better off than those assigned to the treatment group who either failed to enroll or who dropped out in the first three months.

The authors expect that a more positive picture would emerge if information were available regarding the quality of jobs that were held. While queries were made regarding salary levels and probabilities for advancement, too few credible responses were received to show statistically reliable differences between groups.

Conclusions

There is substantial quantitative evidence supporting the success of the Career Intern Program. Considering the number and severity of operational problems the sites encountered, the data are surprisingly good. It is especially noteworthy, however, that when programs were operating smoothly, the results were substantially more positive than when they were experiencing difficulties. The potential benefits to program participants thus appear to be substantially greater than those actually accrued during the demonstration period.

The authors believe that the nature of the demonstration with its extremely tight schedule, unrealistic enrollment quotas, intrusive evaluation, uncertain funding, and other generally negative influences, was responsible for at least some of the difficulties sites encountered. The evidence from at least two of the sites suggests that full and smooth implementation is not an unrealistic expectation, however, given adequate leadership and time for the program to mature. Had all four sites attained this operational status, the results of this evaluation would almost certainly have been substantially more positive.

In conclusion, it is appropriate to reiterate that this report covers only one aspect of RMC's evaluation of the Career Intern Program. The reports of other tasks must also be read in order to obtain a complete perspective on the CIP demonstration. Those reports contain substantial amounts of qualitative data, including several case studies that should be considered in evaluating the program. As is pointed out several times in the main body of this report, these qualitative data lend strong support to the quantitative evidence. Both sources attest to the effectiveness of the Career Intern Program in reshaping the lives of disadvantaged and alienated youths.

I. INTRODUCTION

Background of the Career Intern Program

The Career Intern Program (CIP) is an alternative high school designed to serve disadvantaged and alienated students (called interns) who either dropped out of regular high schools or who were considered potential dropouts. The objectives of the program are to enable students to earn a regular high school diploma (as opposed to a GED), to prepare them for meaningful employment, and to facilitate their transition from school to work. The program offers extensive counseling--academic, personal, and career--and attempts to make academic subjects palatable and relevant to the lives of the students through a heavy infusion of career-oriented content.

Run by a community-based organization, the Career Intern Program enjoys an unusual symbiotic working relationship with the local school district. It serves those students whose needs are not adequately met by the local high school, but the students remain on the local school's books. State monies that are distributed to the schools based on enrollment or attendance thus continue to flow to the local high school even though the students are being served by the CIP. The high schools award diplomas to students graduated by the CIP.

The CIP was initially developed in Philadelphia in the mid-1970s. An independent evaluation conducted by Richard A. Gibboney Associates (Gibboney Associates, 1977) found the program to be successful. The evidence of success was judged sound by the Joint (U.S. Office of Education and National Institute of Education) Dissemination Review Panel, and the program was approved by that group as eligible for federally funded dissemination.

Under authorization of the Youth Employment and Demonstration Projects Act (YEDPA, Public Law 95-93), the Department of Labor (DOL) and the National Institute of Education (NIE) entered into an Interagency Agreement in late 1977 to test the replicability of the CIP and to determine whether the same beneficial outcomes could be obtained in the replication sites. Subsequently, NIE contracted with OIC/A to manage the replication effort. OIC/A then, through a competitive bidding process, selected four local OIC chapters to undertake the CIP replication. Three of the selected sites were urban and one was located in a small (30,000) city.

Overview of the Evaluation

The work statement for the evaluation was prepared jointly by NIE and DOL. Four separate tasks were called for:

- Task A. Conduct studies and analyses as required to answer the questions, "What happens to the Career Intern Program in the process of implementation in additional sites? What factors account for the changes or adaptations, if any? For the fidelity, if any, to the original program goals and practices?" (RFP NIE-R-78-0004, p. 9)

- Task B. Conduct studies and analyses as required to answer the question, "Does the Career Intern Program continue to be effective in helping youth when it is implemented in sites other than the Philadelphia prototype?" (ibid, p. 13)

- Task C. Conduct studies and analyses as required to answer the question, "What happens to young people in the CIP program that could account for its effectiveness?" (ibid, p. 16)

- Task D. Conduct studies and analyses as required to answer the fourth question, "How does the CIP approach compare in effectiveness, feasibility, impact, and factors important for policy with other approaches undergoing comparable evaluations, to helping the population to be served through the Youth Employment Act?" (ibid, p. 20)

To assure comparability with the original CIP evaluation, the work statement specified that the evaluations of the replication sites employ the same instruments and designs as that study. While some modifications were eventually made to strengthen the study, care was taken to preserve the desired comparability.

The present report deals only with Task B. Task A and Task C, however, are highly relevant to the material presented herein as variations in the extent or manner in which individual components of the treatment were implemented almost certainly affected program outcomes. While an attempt has been made throughout this report to relate observed outcomes to implementation events and conditions, small sample sizes (for any particular cohort at any particular site) and other methodological problems place substantial limitations on the extent to which clear-cut relationships can be credibly established. The complexity of implementation events and conditions is another factor which limits the interpretability of outcome findings, and the reader is encouraged to examine the Final Task A Report (Treadway, Stromquist, Fetterman, Foat, & Tallmadge, 1981) and the Final Task C Report (Fetterman, 1981) to gain a fuller appreciation of implementation-outcome relationships.

The CIP replication was originally planned as a two-year demonstration although the possibility of an extension was made known from the outset. It had been anticipated that four cohorts of interns would be enrolled at each of the sites during the original demonstration period. The average size of each cohort was planned to be 75 and at least 2 of the cohorts were to be over-subscribed so that randomly assigned control groups could be formed.

In actuality, only three cohorts were enrolled during the original demonstration period at each of the four sites because of severe recruiting difficulties, and the first two of them were smaller than the 75-member projection. Recruiting difficulties also precluded the formation of control groups for the first and second cohorts at all sites, although control groups were established for the third cohorts at all sites.

A nine-month extension was granted to the four replication sites. During the extension period, fourth cohorts (complete with control groups) were taken in.

The evaluation described herein encompasses the second, third, and fourth cohorts. The first cohort was not included for several reasons:

- the cohort had entered the program at two sites before the evaluation contract was awarded.
- it was felt that the replications needed some time to stabilize and that data collected from the first cohorts would not provide reliable indices of program effects.
- the first cohorts at several of the sites were quite small and it was felt that findings based on such small samples would have been difficult to interpret.

Participating interns and controls were pretested prior to enrollment, midtested sometime between 3 and 6 months after intake (depending on the cohort), and posttested sometime between 9 and 12 months after intake. The test battery consisted of paper-and-pencil tests encompassing reading and math achievement, career awareness, self-concept, and locus of control. Second- and third-cohort interns and third-cohort controls were followed up in the summer of 1980 and again in January/February, 1981. Fourth-cohort interns and controls were followed up only once in January/February, 1981.

Early during the period of recruitment for the third cohort, it appeared that it might not be possible to assemble enough applicants to the program to form both treatment and control groups. For this reason, comparison groups which consisted of (a) low achieving students in the feeder schools, (b) students enrolled in other alternative-school programs, and (c) youths who had dropped out of school were put together to provide alternative baselines against which to measure the success of CIP interns. These groups were mid- and posttested at the same time as the third-cohort treatment and control groups.

Pre-, mid-, and posttest data summaries for all treatment, control, and comparison groups are presented in this report (some were also included in earlier Task B reports). These data were analyzed three different ways, making use of analyses of covariance,

standardized gains, and norm-referenced approaches. The follow-up data were analyzed using (primarily) Chi Square techniques.

Summary of Relevant Implementation Events and Conditions

The most important consideration to keep in mind when reviewing the outcome evaluation findings presented in this report is that the CIP encountered a large number of implementation problems. Many of these problems stemmed either directly or indirectly from the extremely compressed time schedule and bad timing associated with start-up operations. (Contracts were awarded to the replication sites in mid-December, 1977. Staffing and training were accomplished during the remainder of that month and the sites were expected to begin serving students by the end of January, 1978.) A second major source of problems arose from the anxieties felt by both staff and students as the demonstration period drew to an end and futures were uncertain.

These causes underlying implementation difficulties are important because they were functions of the manner in which the demonstration was undertaken and do not necessarily reflect negatively on the transportability of the CIP. Despite the reasons for their existence, however, there can be no doubt that implementation difficulties impacted on the "treatment" that the CIP interns received. In fact, none of the interns in any of the three cohorts studied at any of the four sites experienced twelve months of treatment that was not disrupted by at least one major trauma such as the termination or resignation of the director.

Brief summaries of significant implementation events at each site follow. Subsequent sections of the report refer back to these summaries whenever they appear to be useful in understanding or explaining outcome findings.

Site A. Site A got off to a good start. The director had previous experience in setting up new organizations and proved to be a capable leader in start-up operations. Unfortunately, other key positions were occupied by less well suited individuals. Nevertheless, Site A enrolled its first cohort on March 20th, 1978, and achieved full operational status shortly thereafter.

The second cohort of interns (the first one studied) entered the program on July 24th, 1978. For the following six months (until midtesting in late January-early February, 1979), the program operated relatively smoothly. One problem, however, was that staff turnover was high--11 of 22 staff members left the program, either voluntarily or involuntarily. This high staff turnover served to lower intern attendance rates, yet morale was high among both staff and attending students. There were, however, some significant staffing problems that remained to be solved, particularly in the counseling department.

The third cohort of interns entered the program at the beginning of February, and almost immediately thereafter things began to come apart. The pressures to meet third-cohort enrollment quotas for both treatment and control groups had been intense and had led to interpersonal animosities and a general lowering of morale. Several staff members voiced dissatisfaction with the director's management style which they perceived as authoritarian and unprofessional.

At about this time a staff committee imposed a Code of Conduct and a Dress Code on the interns. The sudden and apparently arbitrary manner in which these new regulations were imposed produced a strong negative reaction on the part of the interns who went so far as to stage a temporary boycott of the program.

More serious problems arose when RMC's first report on implementation was published in March. The negative comments about Site A were carefully culled from that report and related to the CIP staff by the local OIC director without any indication that the report also had many positive things to say. Morale plummeted, dissention rose sharply, and productive program functions ground nearly to a halt. In May the CIP director was forced to resign. A new director was brought in, but recovery was slow.

In late May-early June, the third-cohort interns were midtested. About a month later, second-cohort interns were posttested.

The new director lasted only a few months and was terminated in September. During his tenure, however, there were two other resignations in key management roles.

The counseling supervisor was appointed director in September, 1979. His lack of management experience soon became apparent, however, and the staff reported serious difficulties in communication. Morale did not improve and, in fact, divisiveness among staff members increased. At about this same time, the end of the originally planned demonstration was drawing near. The future of the program was unclear, although there were vague promises of an extension. Staff members began to worry about their future employment and this concern was one more factor that negatively affected program operations and climate.

A nine-month extension was finally granted in December, 1979 and all four sites began feverish recruiting efforts to meet enrollment quotas (90 treatment, 55 control students) by the January 31st, 1980, deadline.

In late January-early February, third-cohort interns were posttested. At about the same time, fourth-cohort interns entered the program. The increased size of the student body improved the climate at Site A temporarily but the widening rift between the director and key staff persons quickly served to offset this gain.

Midtesting of the fourth cohort took place in late May-early June, at the end of the regular school year. As the summer progressed, it became clear that full program funding would not be extended beyond September. Both staff and interns were increasingly concerned about their futures. By the end of August, when the fourth cohort was posttested, the program was in disarray. Intern attendance was well below 50%, three of the top four managers (including the third director) had resigned, and most remaining staff members were new and untrained. Program operations were at a virtual standstill and staff and intern morale were at an all-time low.

Site B. Site B also got off to a good start. Although major problems were experienced in working out an adequate agreement with the LEA, operations ran smoothly once that hurdle had been cleared. The director of the program was well qualified and had strong leadership skills. At least partly as a result of his efforts, capable and caring individuals were found for both counseling and instructional staff positions. Intern recruitment was less of a problem at Site B than at the other sites--and consequently less disruptive of other program functions. The facility, although smaller than would have been desirable, was bright and pleasant and contributed to the overall positive climate of the program.

Site B enrolled its first cohort of interns on April 17th, 1978, before the approval of the LEA had been obtained. The second cohort was enrolled in mid-October. Staff turnover from program start-up until enrollment of the second cohort was limited to two professionals, both of whom had left to take better paying jobs. During the next three months, two math teachers and an aide left the program, also to accept higher paying positions. This pattern of terminations confirmed the fact that the CIP salary scale was not competitive. More importantly for the present discussion, the fact that there were no dismissals and only a few voluntary terminations suggests that hiring practices were unusually effective at Site B and that there was little job dissatisfaction.

As was the case at all other sites, a third cohort of interns was enrolled about the end of January, 1979. The enrollment quotas of 90 interns and 55 controls were met without great difficulty although the entire staff and several interns had to be pressed into recruiting duty.

The large number of interns enrolled at Site B exceeded the housing capacity of the facility and additional space had to be leased in a nearby building. Walking between buildings provided a temptation to "cut out" that some interns found impossible to resist. Attendance fell and the climate at the site suffered somewhat. The counseling staff reported that the large number of interns to be served precluded them from spending as much time with each individual as would have been desirable. Despite these difficulties, the program continued to run smoothly and morale was high among both staff and interns.

Midtesting of second-cohort interns took place in mid-April, 1979, and third-cohort interns were midtested in late May-early June. The program was still operating smoothly but staff morale was beginning to be affected by the low salaries, heavy work load, and, perhaps most importantly, by the lack of vacations comparable to those of teachers in the regular schools.

The second cohort was posttested in mid-October. At that time, and in the course of a mid-November site visit, symptoms of staff burnout were beginning to emerge. This problem was exacerbated by uncertainties regarding the extension of funding beyond the original demonstration period. Intern attendance continued to be somewhat lower than it was before intake of the third cohort, but all aspects of the program continued to be implemented and the climate was generally positive.

In mid-December, the nine-month extension became official. Site B was well prepared and enrolled a fourth cohort in January (a 55-member control group was also formed). Short-term anxieties about the program's future were relieved. Posttesting of the third-cohort interns was accomplished at about this same time.

CIP operations continued much as before until April when the director announced his intention to resign for reasons of career advancement. His resignation had a major impact on all aspects of CIP operations. The deputy OIC executive director was given responsibility for the program when the original director departed. About a month later another OIC person was assigned half time as interim acting CIP director. Unfortunately, the staff perceived these two individuals as temporary employees and behaved accordingly. The lack of leadership took its toll. Intern attendance fell and a number of interns dropped out of the program altogether. Several staff members also chose this time to move on.

Midtesting of fourth-cohort interns was accomplished in late May-early June, 1980. At that time most program components were still functioning smoothly. Much of the enthusiasm observed earlier had disappeared, however, and the morale of both staff and students was low.

By the end of August, when fourth-cohort interns were posttested, attendance was down to about 30% and morale was at an all-time low. While some members of the staff were optimistic that funding would be found which would enable the program to continue, others were actively looking for other employment. Although positive feelings about the CIP continued to be expressed by both the staff and the interns, the program bore hardly any resemblance to what it had been before the original director resigned.

Site C. Site C had a difficult time establishing an acceptable working agreement with the school district. The problem was largely due to pressures brought to bear on the LEA by the local teachers'

union. It was not helped, however, by the fact that the CIP leadership was inexperienced and underqualified. The person appointed director had not, in fact, sought that job. He had applied for the position of counseling supervisor but was named director when no more-qualified person could be found on short notice. Other staff positions were also filled with marginal people and, even after an agreement with the school district had been worked out (through the intervention of OIC/A), staff inadequacies at all levels plagued the program.

Despite the fact that it had no viable working agreement with the LEA, Site C was the first site to enroll interns. This event occurred on February 23rd, 1978. The cohort comprised 38 interns, all of whom had previously dropped out of school.

A working agreement with the LEA was finally signed on July 13th, 1978, but it was not until three months later that a second cohort of interns was enrolled. Severe recruiting difficulties had been encountered and only 46 interns (and no controls) had been signed up. The program, nevertheless, was operating smoothly at the time of RMC's October, 1978, site visit except for the divisiveness and low staff morale that resulted from inadequate leadership.

October, November, and December were months of intensive recruiting activity. Enrollment quotas of 90 interns and 55 controls had to be met by January, 1979, or the program would, most probably, have been shut down. Instructional and counseling activities were reduced to a bare minimum as staff and interns alike engaged in a wide variety of recruiting activities.

During this same time period, the local OIC realized that some action would have to be taken regarding CIP leadership. The OIC executive director temporarily took over the CIP directorship. The original director was retained, however, in the hope that he would learn some of the skills he lacked during the interim period.

The "catchment area" for recruiting was extended to include three additional LEAs. As a result, the enrollment quotas were met. Also as a result, however, the Site C CIP had to accommodate the curriculum and graduation requirements of four LEAs rather than just one. At the time of RMC's second visit to Site C (February, 1979), the entire counseling staff was inundated with paperwork associated with the rostering of the new interns into the courses they needed to graduate. The counselors were frustrated that they had so little time to spend counseling, and their morale suffered as a result.

Confusion regarding CIP leadership also had its impact on staff morale and an atmosphere of paranoia prevailed as various individuals jockeyed for position and maintained written logs of the transgressions of others. The interns, too, sensed the program's disarray. Derogatory graffiti began to appear on the lavatory walls and clusters of students began to "hang out" in the hallways. Even so, they continued to compare the CIP favorably with their former high schools.

On March 2nd, the original director was reinstated on a provisional basis. The situation worsened almost immediately, however, and he was removed permanently at the end of the month.

In April, second-cohort interns were midtested. The program at that time was at its lowest point, but an interim director with appropriate credentials had been appointed and there was some reason for optimism. In early May RMC again visited Site C. While intern absenteeism continued to be high and most program components were being implemented perfunctorally, if at all, staff morale was definitely on the rise.

A strong and well qualified person was appointed permanent director in mid-May. Shortly thereafter, third-cohort interns were midtested.

Gradually the new director began rebuilding the program. Ineffectual staff members were replaced and vacant slots were filled. New procedures were developed and installed and "things began to happen." At about this time, discussions were going on within the Department of Labor regarding the possible extension of the demonstration period. DOL was aware, however, of Site C's problems and was seriously considering extending only the other three sites. Site C knew it was "under the gun." Some staff members were demoralized believing that they would be "sacrificed" to provide an object lesson to the other sites.

A representative of DOL visited Site C in June, 1979, ostensibly to determine whether the site should be terminated. His visit was so perfunctory, however, that site personnel were left with the impression that the decision had already been made. Again morale was negatively affected, but efforts continued to pull the program back together.

The summer was a period of intense revision, reform, and upgrading of operations in preparation for DOL's final review of the program scheduled for October. In September, it became clear that an additional cohort of interns could not be accommodated in the current facility. Feeling that the chances of extension were good (and might be enhanced by a more suitable building), a search was conducted, and a suitable place was located. The CIP moved in October, with staff and students completing the entire moving operation themselves.

On October 30th, 1979, DOL made its long-awaited visit to the site and found it sufficiently improved to be granted the same nine-month extension planned for the other sites. Recruiting activities began in earnest as a goal had been established of enrolling 100 interns and obtaining 75 controls. Posttesting of the second-cohort interns was done at about this time.

In January, 1980, a new cohort of 66 interns was admitted to the program. A control group of 29 members was also found. Although these numbers fell far short of the established quotas, they were accepted. The program, at this time, was almost fully implemented, intern attendance was good, and staff morale was high. While some problems remained, things had never been better at Site C. It was at this juncture that posttesting of the third-cohort interns was accomplished.

RMC visited Site C in April, 1980. Program operations were observed to be running smoothly and intern attendance was high. Staff morale, however, was not as good as during the previous visit. There was evidence of burnout. More importantly, however, the end of the demonstration period was drawing near. Staff were beginning to worry about finding new jobs and complaints about inequitable pay, lack of adequate vacation time, and related issues had begun to surface again. Another contributing factor was that monies promised for the extension period had been held up in Washington. Local funds were used in the interim but they were limited and, for a time, operations proceeded on a day-to-day basis with real concern that the site would have to shut down for lack of funds. Despite these problems, all major program functions were carried out in compliance with the program model.

In late May-early June, fourth-cohort interns were midtested. In mid-June the CIP director announced her plan to resign from the program in mid-August. Subsequently, the instructional supervisor and the reading specialist tendered their resignations. All three left the program in mid-August while RMC visitors were on site. Although the local OIC was confident of its ability to find strong leaders to fill the vacant positions (and there is evidence that they succeeded) morale among the remaining staff members was observed to be very low. During the period between the submission of resignations and actual departures, problems common to lame duck administrations emerged. Staff members who were staying resented those who were leaving and felt disinclined to follow their instructions.

In late August-early September, fourth-cohort interns were posttested.

Site D. Site D, like Site C, did not get off to a good start. The original director not only lacked the skills and experience required by the job, but was guilty of duplicity in dealing with her staff. The local OIC executive director believed in "management by exception" and provided little leadership or guidance.

In mid-April, 1978, the CIP staff moved into the remodeled parochial school that was to house the program for the duration of the demonstration. Both prior to and after that time, the CIP director and the OIC executive director tried unsuccessfully to work out an acceptable cooperative agreement with the school district.

Eventually OIC/A intervened. OIC/A met with the Site D school board on May 5th, 1978, and an agreement was signed five days later.

The first cohort of 23 interns was enrolled at the end of May, 1978. Recruiting for the second cohort began immediately but, by September, it was already clear that Site D would not be able to identify enough candidates to form both treatment and control groups. NIE waived the control group requirement and on October 16th, 41 new interns were enrolled.

RMC visited Site D in November, 1978 and found the program to be in disarray. Most of the problems appeared to be the result of deficient leadership. The director had isolated herself from all of the staff except the instructional supervisor. Most communications to other staff members--even to the counseling supervisor--were by memorandum. Not surprisingly, this situation led to factionalism throughout the remaining staff. Some were deeply resentful and did not hesitate to discuss their feelings with the RMC site visitors. Others chose to side with management, while still others tried to stay out of the conflict and simply do their jobs.

As could be expected, staff morale was low, the program climate was dominated by self-centered concerns, and implementation of instructional and counseling functions was mechanical at best. The interns were sensitive to all of these problems and were attending sporadically. Attendance was observed to be below 50% and, throughout the course of one afternoon, only 9 of the 47 enrolled students were observed in the building.

OIC/A was aware of the worsening situation at Site D and, in December, 1978, prevailed upon the local OIC executive director to remove the CIP director and instructional supervisor. The OIC/A deputy director of the CIP demonstration then stepped in and took control of the program. He remained at Site D for some three months establishing new procedures, training staff, and generally reshaping the program. He also found that relationships with the feeder schools had been impaired by misinformation and negotiated new agreements. Finally, he was instrumental in finding a new director, who joined the CIP on March 12th, 1979.

In January, 1979, while the program was being directed by the OIC/A deputy demonstration director, enough applicants had been recruited to form a third-cohort of interns as well as a control group. At the time of RMC's site visit a month later, staff morale was very high, intern attendance had risen to approximately 70%, and the program climate was positive, caring, and supportive. One of the instructors had been promoted to instructional supervisor and was proving to be both competent and well respected by her staff. While problems remained, the program had improved dramatically and appeared well on its way to full implementation.

The second cohort was midtested in mid-March--just about the time the new director joined the program. She was a strong and

experienced leader who operated with an inclusive and democratic management style. The progress made during the OIC/A intervention continued under her direction. When RMC visited the site again in May, the program was operating smoothly and morale was high among both staff and interns. There had been a substantial number of intern terminations between the March and May visits, but the attrition appeared to be largely the result of excessively zealous recruiting. Interns had been taken into the program in order to meet enrollment quotas who were not adequately motivated and who never seriously intended to remain. It was at about this time (mid-May, 1979) that third-cohort interns were midtested.

Over the summer of 1979, the CIP ran a reduced program to accommodate the interns' need for employment. Arrangements were made with several summer youth programs that enabled interns to attend classes in the mornings and work in the afternoons. In September, the CIP resumed full operations when the public schools reopened. In mid-October the second-cohort interns were posttested.

RMC visited Site D again in December, after the program had been granted an extension through September of 1980. About 65 interns from the first three cohorts were still active and, although the numbers were small, staff and student morale were high; the program climate was very positive, and program functions were operating very well. Recruiting for the fourth cohort was underway and it was clear that there would be little difficulty in meeting enrollment quotas. Relationships with the feeder schools had become so positive under the new director's leadership that the CIP was allowed to set up recruiting booths in the buildings and use the public address system for announcements.

In January, 1980, a new cohort of 100 interns was enrolled. At approximately the same time, third-cohort interns were posttested. RMC visited the site again in March and found the program still running smoothly. Attendance had stabilized at about 70% and morale continued to be high. Staff turnover (mostly f · reasons of advancement) was somewhat of a problem, but the program seemed able to attract well qualified replacements for those who left.

Plans were underway for obtaining funds from alternative sources so that the program could continue beyond the nine-month extension period. Proposals had been submitted to the CETA prime sponsor, a private foundation, and the state. Everyone was optimistic about the outcomes and there was little of the concern over job security that was observed at the other sites.

In May, 1980, fourth-cohort interns were midtested.

RMC's final visit to Site D occurred in August, 1980, while fourth-cohort interns were being posttested. The situation was much as it had been in March. The program was operating smoothly and both interns and staff were enthusiastic and working hard. One of the long-term staff members commented, "It's smooth sailing now," as she recalled her first year and a half with the CIP.

II. METHODOLOGY

This study has employed a variety of data analyses techniques. The majority of these techniques were applied to the analysis of scores on paper-and-pencil tests administered to CIP interns and members of the control and comparison groups prior to the intake of each cohort at each site and approximately 6 and 12 months thereafter. These analyses are discussed immediately below. Follow-up data were also collected approximately 6 and 12 months after post-testing. For the most part, these data consisted simply of frequency counts of youths in various school and employment categories. The methods used to analyze these data are discussed at the end of this chapter.

Analyses of Test Data

As mentioned in the Introduction, this portion of the study encompassed the simultaneous implementation of a control group experimental design, a comparison group design, and a norm-referenced design.[1] Only the control group design was called for in the request for proposal, but a decision was made by the time of contract award to supplement it with a norm-referenced evaluation since large and (possibly) differential attrition of students was expected from the treatment (CIP) and control groups. Such attrition, if it occurred, could create serious doubts regarding the validity of inferences drawn from comparisons between treatment and control groups.

The evaluation was further supplemented by the inclusion of various comparison groups approximately nine months after the study began. This step was taken because the sites were experiencing serious difficulties in recruiting sufficient numbers of students to fill treatment group quotas while also providing adequate numbers for the control groups. It was feared that control groups might have to be abandoned altogether or that they would be too small to provide a stable baseline against which to measure treatment effects.

Constraints were imposed on the evaluation by a number of circumstances associated with CIP operations at the four sites. These constraints typically required that the standard procedures associated with each design be modified. In some cases the modifications were substantial and significantly affect the manner in which the analyses should be interpreted. While the authors believe

[1] All of the various designs that were used attempt to measure the impact of the CIP. Each, however, rests on different sets of assumptions and asks a slightly different question. Appendix A presents a comparison of the designs in these terms. It is included for the methodologically inclined reader and need be of no concern to others.

that the inferences they have made and the conclusions they have drawn are sound and credible, the reader is advised to note carefully all the cautions and caveats contained in the following descriptions of how each design was implemented.

Instrumentation

The study described herein used much the same instrumentation as was used in the evaluation of the original CIP in Philadelphia (Gibboney Associates, 1977). Both evaluations used standardized reading and mathematics achievement tests. The original study used subtests of the Stanford Achievement Test (1973 edition) while the present study used the Metropolitan Achievement Test (1978 edition) because the latter instrument was considered to be substantially better-suited for use with the CIP target population than the former. Before the final selection was made, a careful examination of thirteen of the most commonly used achievement tests was undertaken. A summary of this evaluation is included as Appendix B of this report.

Other instruments used in the original study were the Career Development Inventory (Super, 1970), the Self-Esteem Inventory (Coopersmith, 1967), the Internal-External Scale (Rotter, 1966), and the Standard Progressive Matrices (Raven, 1940). These same instruments were used in the present study (except for the Standard Progressive Matrices, copies are included in Appendix C. There was one difference, however; the Standard Progressive Matrices test was used both pre and post in the original study whereas it was used only as a pretest in the present study. This change was made in response to a suggestion made by the NIE Project Officer.

With the exception of several pretest sessions at one site, all testing was accomplished by RMC-employed site assistants with appropriate professional qualifications. The few test sessions not conducted by RMC were run by a senior-level graduate student in psychometrics who was employed as a CIP math teacher at the time. He was trained by the regular RMC tester at that site and was judged to be well qualified.

The Control Group Design

The evaluation of the original CIP in Philadelphia made use of a randomly assigned control group in order to generate a baseline against which the growth of CIP participants could be measured. More candidates were recruited for the program than could be served, and a lottery-like procedure was then used to determine which applicants would be assigned to the control group and which would be admitted to the program. At mid- and posttesting times, members of the control group were paid to complete the instruments.

The evaluators noted several problems with this approach (Gibboney Associates, 1977). First, many of the control group students who returned for mid- and posttesting lacked motivation

and were observed to mark their answer sheets at random. While an attempt was made to compensate for this problem through application of a statistical adjustment, the results were unsatisfactory. (See Appendix C for a discussion of valid and invalid uses of the correction for guessing.)

A second problem was that attrition from both treatment and control groups was very high (approximately 50%) at the time of mid-testing and 70% by posttest time (see Table 2, p. 28 for a breakdown by site and by cohort) and it seemed likely that attrition from the two groups was non-random and that biases might have resulted which would compromise inferences drawn from subsequent treatment-control comparisons. The more able or more highly motivated control group students, for example, might have changed schools or taken jobs thus making them unavailable for mid- or posttest data-collection sessions. On the other hand, the treatment group students who were not present for these sessions could easily have been those at the other extreme of the distribution who would or could not do the work required to remain in the program. Other hypotheses may be equally plausible, but the fact remains that while random assignment may have assured parity between the original groups, that parity could well have been destroyed by differential attrition.

RMC attempted to deal with each of these problems through design modifications. To control for random responding, students were paid for correct responses on those instruments where responses could be judged either correct or incorrect. The details of this incentive payment strategy are discussed below.

To combat the differential attrition problem, a decision was made to adopt a matching strategy that entailed the formation of dyads or triads of students (depending on treatment and control group quotas) who were as much alike as possible in terms of identifiable, educationally relevant characteristics. One member of each dyad or triad was then selected for the control group while the remaining members were invited to enroll in the CIP. The plan was to limit comparisons between treatment and control groups to those dyads or triads where it was possible to obtain mid- and/or posttest data on the control group member and at least one treatment group member. While this procedure would reduce the size of the evaluation sample, it would also presumably eliminate the bias that might otherwise have resulted from differential attrition. The details of the matching procedure are also described below.

Incentive payment strategy. Because it seemed likely that students with no stake in the study (control and comparison group students) would not put forth their best efforts when responding to mid- and posttest questions, a decision was made to provide an incentive, in the form of a cash payment, for correct responses. Thus, in addition to paying students $10.00 for coming to data-collection sessions, they were paid $.07 for each item they answered correctly. To avoid the problems that might have arisen from

differential reinforcement, members of the treatment group received the same cash incentives.

There were 190 correct answers on the reading and math achievement tests, the Standard Progressive Matrices, and the Information scale of the Career Development Inventory (items making up the other instruments or scales had no correct answers). Students could thus earn as much as $13.30 for correct responding plus the $10.00 for attending. In fact, typical payments were in the $18.00-$20.00 range. Tests were scored immediately, and students were paid in cash or by check within minutes of completing the last instrument.

While the incentives were considered generous, it became clear that they were not entirely successful in achieving the desired results. In one instance, students who had been scheduled for testing were observed playing basketball on a court outside the school. They could not be lured in for data collection. In at least two other instances, students were observed marking their answer sheets without referring to the test booklet. Despite these occurrences, it seemed clear that the incentive strategy was at least moderately successful. The majority of test scores appeared to be valid and the anomalies observed in the Philadelphia evaluation data (e.g., mean posttest scores being lower than mean pretest scores) were eliminated.

Incentive payments were made to members of the comparison groups at pre-, mid-, and posttesting times. There were no such payments to treatment or control group members at pretest times since they were motivated to do well in order to qualify for admission to the CIP. Both treatment and control group members were paid, however, at mid- and posttest times. While treatment group members would probably have been adequately motivated without incentive payments, there was evidence that they would have resented not being treated in the same manner as the other groups.

Matching treatment and control students. The variables on which students were matched were primarily pretest scores and age. Separate matchings were undertaken for reading and math. Where a surplus of good matches could be achieved on the two primary variables, grade level, and number of academic credits needed to graduate from high school were also considered. This set of criteria was incomplete and would have been expanded to include at least pre-CIP school attendance rates had it been possible to obtain this information. Nevertheless, a large proportion of total among-student variance was brought under experimental control by the matching process.

It was not expected that perfect matches could be achieved even under ideal circumstances. As it happened, however, circumstances were far from ideal. Severe problems were encountered in recruiting adequate numbers of students to meet treatment group quotas. For this reason there was no control group for the second cohort and the plan to serve four cohorts during the original demonstration periods

had to be abandoned (although a fourth cohort was served during the extension period).

Recruitment for the third and fourth cohorts extended over a very long time period. Many pretesting sessions had to be scheduled with small numbers of candidates tested at each session. Program staff at the CIP sites felt that potential interns were being lost due to lengthy delays between being tested and being informed as to whether or not they would be admitted to the program. As a result, they requested that treatment and control group assignments be made at the end of each week in which testing occurred and that candidates be notified of their status.

The need to assign students to treatment and control groups on a weekly basis interfered substantially with the matching process. Typically, data were available on only a few students, and the formation of well matched dyads or triads was often impossible. Despite this difficulty, the matching procedure was continued (as well as it could be) and selection of students for the control group continued to be random from each dyad or triad. It was felt that, while treatment and control group assignments could not be changed, it would be legitimate to improve the matching of treatment with control group members after all the students had been pretested (Cook & Campbell, 1979, pp. 47, 48). Such post-hoc matching, of course, would have to be done without any knowledge about the status of students after the pretest since such knowledge (e.g., that a student selected for the treatment group had chosen not to enroll) could clearly bias the matching process and, thereby, the results of any subsequent analyses.

The matching (or rematching) process was further complicated by the fact that pretesting spanned a time interval of more than four months. Because reading and math skills develop over time, it seemed unlikely that a student would obtain the same test score if tested in late January that he or she had actually obtained when tested in the middle of the preceding September. It follows that two students who obtained identical scores tested at widely different times would not have obtained identical test scores had they been tested at the same time.

Adjusting test scores for different testing times. Because of the problem just discussed, it was considered necessary to attempt some form of statistical adjustment to obtain estimates of the scores students would have achieved had they all been tested at the same time. This adjustment was accomplished for reading and math achievement-test scores through use of normative data. The procedure was as follows.

The assumption was made that students whose scores placed them at a particular percentile rank in the national distribution at time T_1 would tend to score at the same percentile rank at time T_2. (This same equipercentile assumption also underlies the norm-referenced evaluation design described later in this chapter.)

Given the equipercentile assumption, a test score, and a test date, it follows that interpolating between adjacent empirical normative data points can yield estimates of the score that would have been obtained on any other particular test date. Unfortunately, the process is not quite as clear-cut as it appears on the surface.

The most salient complication to the interpolation process stemmed from the fact that percentiles do not constitute an equal-interval scale. Thus, if a test score obtained half-way between adjacent empirical normative data points was found to correspond to the 25th percentile in the earlier norms and the 5th percentile in the later norms, it would be incorrect to infer that the interpolated value would be the 15th percentile. (The 12th percentile actually lies midway between the 25th and the 5th.) This particular difficulty was overcome by converting percentiles to normal curve equivalents (NCEs)[2] before interpolating.

The second complication related to the fact that cognitive growth rates are not linear over the twelve months of each calendar year. This complication could not be resolved as satisfactorily as the first because little is known about the exact shape of the growth function. What is known, however, is that growth is slower over the summer than during the school year--particularly for low-achieving students (Tallmadge, 1978; National Institute of Education, 1978; Thomas & Pelavin, 1976; Tallmadge & Horst, 1976). This difference in growth rates can easily be seen in most test publishers' norms tables by comparing the gain in standard-score points per month between fall and the following spring with the gain between spring and the following fall. Unfortunately, it seems likely that further non-linearities exist since the spring-to-fall interval usually ranges from sometime in April to sometime in October and thus encompasses several months of the school year as well as the summer vacation.

If one assumes that cognitive growth proceeds at one more-or-less-constant rate while school is in session, and at a slower, but also constant rate over the summer, then it would be appropriate to use the October-to-April growth rate from September to June and, subsequently, to determine a June-to-September growth rate using whatever annual gain remains. Although alternative rationales could have been developed (e.g., it could have been assumed that start-up would be slow and that the school year would end with a tailing off of growth), the approach described was the one adopted.

---

[2] Normal curve equivalents are normalized standard scores with a mean of 50 and a standard deviation of 21.06 (when a nationally representative sample of any age/grade group is tested). They match percentiles at values of 1, 50, and 99 but, under the assumption that the attribute measured is normally distributed in the population, they constitute an equal-interval scale.

September 15th and June 15th raw-score-to-NCE norms tables were generated by extrapolating from the October 15th and April 20th Metropolitan Achievement Test normative data points. These extrapolated norms tables were subsequently used to obtain interpolated NCEs for each student as a function of his/her own particular testing date.

Caution. A word of caution should be inserted at this point. The procedure just described must be regarded as a poor substitute for testing all students on the same date. (For norm-referenced evaluations, the testing date should also correspond to one of the test's empirical normative data points.) While the authors believe the approach taken was sound--and that there was no better way to deal with the need for staggered testing--small errors have almost certainly been introduced. It seems unlikely that the magnitude of such errors would be sufficient to obscure any educationally significant treatment effect, but even that possiblity must be acknowledged.

Selecting appropriate norm groups. An additional problem needs to be mentioned. Most CIP interns ranged from 16 to 21 years of age but a few exceptions to this age-range requirement were made for various reasons. Many program participants were dropouts who had been out of school for varying amounts of time. Most of those who had not dropped out were classified as juniors or seniors in their respective high schools even though they lacked too many credits to graduate with their classes. Others had been held back one or more years. For these various reasons, it was often not clear what norms tables were most appropriate for individual students.

Ultimately, a decision was made to categorize students according to their ages rather than their grade levels. The age of each student as of October 2nd of the academic year they entered the program was determined. Youths whose ages were between 14 and 14.95 were treated as 9th graders. Those between 15 and 15.95 were treated as 10th graders. Those above 16 were treated as 11th graders. Regardless of their ages, no students were treated as 12th graders at pretest time since 12th-grade norms (the highest level of norms tables) had to be reserved for use with the posttest scores of interns classified as 11th graders when they entered the program.

Out-of-level testing. A final but minor problem related to the test-norming issue is that all treatment, control, and comparison group students were tested out of level. That is, although the majority of the students could be considered as 10th, 11th, or 12th graders, they were tested with the level of the Metropolitan Achievement Test (Advanced Level 1) intended for 7th-through-9th graders. This testing approach was adopted deliberately in view of the fact that most of the students tested were known to be low achievers. Many would find the in-level test too difficult, and their scores, as a result, would be unreliable.

Although the test itself was designed for students in grades 7, 8, and 9, it was possible to gain access to 10th-, 11th-, and 12th-grade norms by means of the (vertical) scale scores. With the Metropolitan Achievement Test (1978 edition) the process is as follows: (a) the out-of-level raw score is converted to a scale score, (b) the scale score is converted to an in-level percentile rank, and (c) the in-level percentile rank is converted to an NCE.

Unadjusted measures. The techniques used to adjust achievement-test scores for differences in testing dates could not be applied to the Career Development Inventory, the Internal-External Scale, or the Self-Esteem Inventory because no normative data were available. Since none of these measures was used for matching, however, and since the ratio of treatment to control group students was approximately the same for each testing date, no biases in treatment-control analyses should have resulted from this failure to adjust.

Systematic influences may be present in the treatment-vs.-comparison-group analyses since most comparison-group students were tested later in the year than treatment and control students. On the other hand, the nature of the measures, coupled with the fact that the treatment did not begin until after all students (treatment, control, and comparison) had been pretested suggest to the authors that the differences in testing times would not significantly affect the evaluation findings. Again, however, readers are cautioned that this inference may be questionable.

Analyzing the data. It was originally intended that all treatment-control comparisons would be based on intact, matched dyads or triads of students. This strategy was employed to counteract the potentially biasing influences of differential attrition. Unfortunately, the rate of attrition was very high and the number of intact groups available for analysis was correspondingly low at all sites. Matched-groups analyses were undertaken, but they were supplemented with covariance and standardized-gain analyses in order to capitalize on the larger sample sizes that were available for these analyses.

The matched-groups analyses were all performed using *t* tests for paired observations. This type of analysis is exactly comparable to a single classification analysis of variance. These analyses were done separately for each site and for each criterion variable.

The covariance and standardized-gain analyses employed in this study were conducted using three somewhat different approaches. Traditional covariance analysis (see Winer, 1971) employs a common, within-group (treatment and control) post-on-pretest regression line. Similarly, the traditional standardized-gain analysis makes use of a common, within-group principal axis of the treatment and control groups' bivariate scatter plots (see Kenny, 1975, and

Tallmadge, 1978). In both cases, the underlying assumption is that these within-group statistics provide better estimates of the population values than either of the individual lines. Because this assumption may often be questionable, RMC elected to conduct three versions of each analysis, one using the control group's regression line/principal axis, one using the treatment group's, and one using the common within-group regression line/principal axis. Interpretations of these analyses are given in the Results section of this report.

The traditional covariance analyses (that used the common, within-group regression line) employed the standard F test (Winer, 1971, p. 772). Exact F tests were not worked out for the covariance analyses that employed just the comparison group's regression line or just the treatment group's regression line. Approximate F ratios were calculated using the denominator from the standard covariance analysis and the gain estimate squared as the numerator (gain = treatment group's adjusted mean posttest score minus comparison group's adjusted mean posttest score). Although the values calculated in this manner may differ slightly from the exact, least squares Fs, the differences should be small in all cases and should not affect any interpretations of the results.

To the authors' knowledge, no exact F test has yet been worked out for standardized-gain analysis. The approximation used here was

$$F = \frac{(\text{Difference in adjusted posttest scores})^2}{\frac{SS_T + SS_C}{n_T + n_C - 2}\left(\frac{1}{n_T} + \frac{1}{n_C}\right)}$$

where

$$SS_T = \Sigma Y_T^2 - (\Sigma Y_T)^2/n_T + b^2[\Sigma X_T^2 - (\Sigma X_T)^2/n_T] - 2b[\Sigma X_T Y_T - (\Sigma X_T)(\Sigma Y_T)/n_T]$$

$$SS_C = \Sigma Y_C^2 - (\Sigma Y_C)^2/n_C + b^2[\Sigma X_C^2 - (\Sigma X_C)^2/n_C] - 2b[\Sigma X_C Y_C - (\Sigma X_C)(\Sigma Y_C)/n_C]$$

and

$$b = \frac{(\bar{Y}_T - \bar{Y}_C) - (\hat{\bar{Y}}_T - \hat{\bar{Y}}_C)}{(\bar{X}_T - \bar{X}_C)}$$

As was the case with the covariance analyses, three different versions of the standardized gain analysis were computed, one using the slope of the common, within-group principal axis, one using the slope of the comparison group's principal axis, and one using the slope of the treatment group's principal axis.

The Comparison Group Design

Approximately nine months after this study began, recruiting difficulties experienced at all four sites made it clear that control groups available for the study would be of minimally acceptable size. For this reason DOL/NIE decided to supplement the evaluation through the employment of various comparison groups.

A brief feasibility study led to the conclusion that, in three of the sites, it would be possible to form comparison groups of (a) potential dropouts in feeder high schools who had not applied for admission to the CIP, and (b) participants in other alternative-school programs. In one of these three sites it appeared that a group of actual dropouts not participating in any academic program could also be assembled. The future of the fourth site (Site D) was uncertain at that time; therefore no attempts were made to form comparison groups.

Most members of the various comparison groups were pretested in January, 1979. A few were tested in late December, 1978, and a few in early February, 1979. They were mid-tested in May and June, 1979, and were posttested in February and March, 1980. Raw scores on the reading and math achievement tests were converted to interpolated NCEs using the same procedures employed with the treatment and control groups. No adjustments were made to scores on the other instruments to compensate for differences in treatment and comparison group testing dates. Unfortunately, these differences are more likely to impact on the comparison group analyses than on the control group analyses. While students were assigned to treatment and control groups shortly after each pretesting session, thereby effecting a proportional balance, this was not the case with the comparison group. All comparison group students were pretested near the end of the four-month interval during which treatment group students were pretested.

All comparison group analyses were done using covariance and standardized-gain procedures. Pretest scores were used as the single covariate.

The Norm-Referenced Design

Norm-referenced evaluations of various types have been popular for many years. Recently one such design was developed for nationwide use in evaluating projects funded under Title I of the Elementary and Secondary Education Act (Tallmadge & Wood, 1976). Evidence from a study which compared gain estimates derived from that norm-referenced design with ones derived from simultaneously implemented,

random-assignment experiments, suggests that the two types of estimates are about equally accurate--at least under the circumstances that were studied (Tallmadge, 1981).

The model is based on what has come to be known as the equipercentile assumption that was referred to earlier. This assumption holds that, in the absence of any special educational intervention, students will retain their percentile (or NCE) status with respect to a norm group over time. Pretest status thus becomes predicted posttest status, and gains are measured by subtracting predicted posttest status from actual posttest status (Posttest NCE - Pretest NCE).

There are two steps in the procedure recommended for implementing the norm-referenced model that were not feasible in the CIP evaluation. First, all testing (pre-, mid-, and post-) should be accomplished within about two weeks of the test's empirical norming date(s). Unfortunately, not only did the cohort intake dates preclude such timing, but recruiting difficulties necessitated extending the pretesting period over four months (in the case of the third cohort). In an attempt to deal as effectively as possible with this problem (as mentioned earlier), the Metropolitan Achievement Test's October 15th and April 20th norms were first extrapolated to September 15th and June 15th. Each student's raw score was then converted to an NCE by interpolating between the extrapolated norms tables according to his or her individual testing date. Some error was certainly introduced by this procedure, but its magnitude is thought to be small and cannot be accurately predicted.

The second model-implementation problem concerned the rule that a single set of test scores cannot be used both to select students for participation in a program and as their pretest measure. When this rule is violated, a spurious regression to the mean occurs, and gains are artifactually either inflated or reduced. In the CIP, students were required to read at the fifth-grade level (more accurately, the entry criterion was set at one standard error of measurement below the fifth-grade reading level). Some candidates scored below this level and were denied admission to the program. To the extent that this happened, students were indeed "selected on the pretest," since they were not re-pretested after being accepted into the CIP.

In the authors' opinion, the biasing influence of pretest selection was small because, except in one site, the great majority of students scored well above the cutoff. To the extent that a bias does exist, however, it will cause gain estimates to be too low. The norm-referenced evaluations will thus tend to be conservative. Real gains may be slightly higher than the norm-referenced estimate.

All norm-referenced evaluations were conducted using the standard paired-observations t test.

Analyses of Follow-Up Data

The test-score analyses described above involved all control group students who could be attracted to the data collection sessions by the monetary and other incentives that were offered. As far as the treatment group was concerned, only those interns who were active participants at the time of testing or who had graduated were included. No attempt was made to test youths who had been invited to join the program but who had failed to enroll or who had terminated prior to the testing session. For the two follow-up studies that were undertaken (the first in the summer of 1980 and the second in January/February, 1981), a slightly different approach was taken.

Attempts were made to contact all youths assigned to the treatment groups and all assigned to the control groups. If direct contact could not be established, information about these youths was sought from school personnel and records; and from relatives, friends, and neighbors. Had we succeeded in obtaining information on all of the youths, the "true experiment" with which the study began would have been preserved. Whatever "treatment effects" might have emerged from the analyses would have been unaffected by possible self-selection biases and highly credible. Despite intensive efforts that included door-to-door canvassing of neighborhoods, however, the return rate was slightly below 80%. (See Table 53, p. 88, for a breakdown by site and by cohort.) While this return rate was surprisingly high considering the much smaller number of youths from whom it was possible to obtain test scores, it was not high enough to remove all possibility of bias resulting from differential attrition. Still, it should have reduced it.

While including untreated members of the treatment group in the analyses serves to maintain the integrity of the design, it also minimizes the size of treatment effect estimates, since gaines made by treated students are at least partially offset by the zero expected gains of the untreated students. The latter consideration led RMC to subdivide the treatment group into treated (those who enrolled in the CIP and remained a minimum of three months) and untreated (those who did not enroll or left the program in less than three months) subgroups. In weighing evidence from the two follow-ups, it should be kept in mind that comparisons between treatment and control groups will systematically underestimate the size of treatment effects while those between the treated subgroup and either the untreated subgroup or the control group will systematically overestimate treatment effects (because of self-selection bias).

The follow-up data lent themselves to two major comparisons. The first compared groups in terms of high school status. The proportions from each group who had graduated from high school, were currently enrolled, or had earned GEDs were contrasted with the proportion who had dropped out of school prior to graduation and had not earned GEDs. The second major comparison contrasted groups

in terms of those members who were either enrolled in school (high school, college, GED, or vocational) or employed, as opposed to those who were neither enrolled nor employed.

Data from the first and second follow-ups were analyzed separately. The analyses were conducted separately by site and by cohort as well as across cohorts and across sites, just as was done with the test score analyses.

III. RESULTS

This chapter summarizes the findings of the entire outcome evaluation task. It is organized under three major headings: (a) holding power, (b) test-score outcomes, and (c) follow-up findings.

Holding Power

Table 1 presents the numbers of treatment and control group youths who were: (a) pretested, (b) midtested, and (c) posttested by site and by cohort. Table 2 presents the same data but reduced to attrition rates from pre-to-midtest and from pre-to-posttest. These data are intended to provide some indication of the CIP's ability to retain youths after they enrolled. Unfortunately, for reasons explained below they are somewhat misleading.

Table 1
Sample Sizes by Site and Cohort
at the Time of Each Testing

		Pretest		Midtest		Posttest	
Site	Cohort	Treatment	Control	Treatment	Control	Treatment	Control
A	II	65	--	21	--	18	--
	III	108	55	32	19	22	16
	IV	101	55	30	27	21	18
	Total	274	110	83	46	61	34
B	II	76	--	40	--	15	--
	III	121	60	88	25	50	20
	IV	75	74	41	32	32	26
	Total	272	134	169	57	97	46
C	II	49	--	28	--	9	--
	III	120	54	47	30	21	14
	IV	66	29	53	10	34	12
	Total	235	83	128	40	64	26
D	II	67	--	15	--	6	--
	III	118	55	52	15	33	16
	IV	176	106	77	54	67	50
	Total	361	161	144	69	106	66
All	II	257	--	104	--	48	--
	III	467	224	219	89	126	66
	IV	418	264	201	123	154	106
	Total	1142	488	524	212	328	172

Table 2
Attrition Rates by Site and Cohort

Site	Cohort	Pre- to Midtest		Pre- to Posttest	
		Treatment	Control	Treatment	Control
A	II	68%	--	72%	--
	III	70%	65%	80%	71%
	IV	70%	51%	79%	67%
	Combined	70%	58%	78%	69%
B	II	47%	--	80%	--
	III	27%	58%	59%	67%
	IV	45%	57%	57%	65%
	Combined	38%	57%	64%	66%
C	II	43%	--	82%	--
	III	61%	44%	82%	74%
	IV	20%	66%	48%	59%
	Combined	46%	52%	73%	69%
D	II	78%	--	91%	--
	III	56%	73%	72%	71%
	IV	56%	49%	62%	53%
	Combined	60%	57%	71%	59%
All	II	60%	--	81%	--
	III	53%	60%	73%	71%
	IV	52%	53%	63%	60%
	Combined	54%	57%	71%	65%

As can be seen (most easily from Table 2), the attrition rates from the treatment and control groups are quite similar when computed across sites for cohorts III and IV as well as across the three cohorts. None of the differences even approaches statistical significance. This finding appears to suggest that the program's ability to retain youths was quite low. This appearance, however, is very deceiving. All youths assigned to the control group were encouraged (and paid) to participate in the mid- and posttesting sessions. Of those assigned to the treatment group, however, only those still enrolled in the program and those who had graduated were permitted to take the mid- and posttests. In other words, youths had to do something to stay in the treatment group but nothing to stay in the control group.

If one makes the assumption that some of the ineligible members of the treatment group would have returned for testing had they been invited, a very different picture emerges. To illustrate, suppose ineligible treatment group members would have returned for testing at half the rate at which members of the control group returned (a conservative estimate, we believe). Had this happened, there would

have been 48 more third-cohort treatment group members at midtest time and 49 more at posttest time. The corresponding increases for the fourth cohort would have been 51 and 53.

From pre-to-midtest the attrition rate for the third-cohort treatment group would thus have been 43% compared to 60% for the control group. This difference would have been significant at the .01 level, one tailed (Chi Square = 6.17, df = 1). Pre-to-midtest attrition rates for the fourth cohort would have been 40% for the treatment group versus 53% for the control group. This difference would also have been statistically significant (Chi Square = 3.64, df = 1, $p < .05$ one tailed).

Pre-to-posttest attrition rates would not have been significantly lower for either the third- or fourth-cohort treatment groups than for the corresponding control groups. If the two cohorts were combined, however, the treatment group rate (57%) would then have been lower than that of the control group (65%) at the .05 (one tailed) confidence level (Chi Square = 3.54, df = 1).

It is not clear exactly how these numbers should be interpreted. It does seem, however, that they provide reasonably convincing evidence of the existence of a treatment effect.

A literal interpretation can say only that significantly higher percentages of treatment group members could have been mid- and posttested than was the case for members of the control group. However, since this difference is clearly attributable to those members of the treatment group who attended the program (non-attending treatment group members were assumed to return for testing at a rate only half that observed in the control group), a case can be made that the program did have significant holding power.

At this juncture, it should be pointed out that treatment group students who were attending the program were easier to locate and inform of the testing sessions than control students. This difference no doubt contributed somewhat to the apparent treatment effect. The authors do not believe, however, that it could have been totally responsible.

Between-site and between-cohort differences are somewhat easier to interpret. Although the across-cohort, pre-to-posttest (unadjusted) attrition rates among treatment group members are not significantly different over all sites (Chi Square = 6.90, df = 3, $.10 > p > .05$ two tailed), a comparison of Site A with Site B produced a significant Chi Square (6.58 with 1 degree of freedom, $p < .02$ two tailed). The direction of the difference, furthermore, is consistent with the general impression that Site A had the least success in attaining full program implementation, while Site B was fully implemented for the largest portion of the demonstration period (see site descriptions in Chapter I).

Perhaps the biggest difference between sites occurred at the beginning of the demonstration period when Sites A and B got off to good starts while Sites C and D suffered through serious management problems. Second-cohort pre-to-posttest attrition rates for treatment group youths are again consistent with this observation, being lower at Sites A and B (77%) and higher at Sites C and D (87%). The difference between these attrition rates is statistically significant (Chi Square = 3.18, df = 1, $p < .05$ one tailed).

During the tenure of the third cohort, Sites B and D were functioning well while Sites A and C continued to experience difficulties. Third-cohort attrition from Sites B and D was 65% while that from Sites A and C was 81%. Again, this difference is highly significant (Chi Square = 8.64, $p < .01$).

While the fourth cohort was attending, Site D attained full implementation while Site A continued to have a difficult time. Attrition rates for the two sites were 62% and 72% respectively. This difference, too, is statistically significant at the .05 level (Chi Square = 4.76). Sites B and C continued to operate well, at least for a large percentage of the time, despite the resignations of their directors. A comparison of the combined attrition rates for Sites B, C, and D (58%) with that observed at Site A is again highly significant (Chi Square = 7.43, $p < .01$).

Program implementation deteriorated at Site A as a function of time while it improved at Sites C and D. The improvement at Site D occurred earlier, however, than at Site C. To determine whether these implementation changes were accompanied by corresponding changes in attrition rates, the following comparisons were made:

- At Site A, the pre-to-posttest attrition rate of the second-cohort treatment group was compared against that of the third- and fourth-cohort treatment groups combined. The difference, while in the predicted direction, was found not to be statistically significant (Chi Square = .93).

- At Site C, the pre-to-posttest attrition rate of the second- and third-cohort treatment groups (combined) was compared against that of the fourth-cohort treatment group. The difference was in the predicted directions and statistically significant at the .001 level (Chi Square = 14.17, df = 1).

- At Site D, the pre-to-posttest attrition rate of the second-cohort treatment group was compared against that of the third- and fourth-cohort treatment groups combined. Again, the difference was in the predicted direction and was statistically significant at the .005 level, one tailed (Chi Square = 10.34, df = 1).

These various findings, taken together, constitute a convincing body of evidence that attrition is inversely related to the quality

or extent of program implementation. When the CIP is well implemented it has significantly better holding power over participating students than when it is less well implemented.

Test Score Outcomes

The following pages contain a complete summary of all analyses performed on test scores during the three-year CIP demonstration period. These analyses are organized first by subject matter and then by type of analysis within subject matter. Finally, for each type of analysis, the pre-to-midtest results are presented prior to the pre-to-posttest results.

Reading

Tables 3 and 4 present the results of the norm-referenced analyses performed on treatment groups scores. Table 3 summarizes the findings of the pre-to-midtest analyses while Table 4 encompasses the pre-to-posttest findings.[3] As can be seen, most of the gains are statistically significant. Combined across sites and cohorts, the mean pre-to-posttest gain is 2.6 NCEs and the pre-to-posttest gain is 6.7 NCEs (just short of one-third of a national-sample standard deviation).

The pre-to-midtest results, when combined across sites, show that the smallest gain was made by fourth-cohort students. This finding is perhaps best explained by the short pre-to-midtest interval for the fourth-cohort (3.5 to 4 months). The largest gain was made by third-cohort students--a fact largely attributable to the results at Site D. While the large gain at Site D may have resulted from the dramatic turn-around that occurred at that site, the small negative gain made by fourth-cohort students, when implementation at Site D was even better, seems to contradict this hypothesis. It may be that the disruption which followed enrollment of the large fourth cohort (130 interns) was responsible for the poor showing, but that inference borders on pure speculation.

---

[3] The analyses reported here all employ *t* or *F* tests. Because many such tests are reported, their tabled probability levels are too low. While this problem could theoretically have been avoided by employing one overall analysis of variance and various subanalyses within it, the design would have been extremely complex. Furthermore, interpretive explanations of results at the level of fourth-order interactions (where individual site, individual cohort, single criterion, norm-referenced evaluations would fall) are so cumbersome that the distorted probability levels of multiple *t* and *F* tests were viewed as the lesser of two evils.

Table 3
Treatment Group Pre-to-Midtest NCE Gains in Reading:
Estimates Derived from Norm-Referenced Analyses

Site	Cohort	Pretest NCE Mean	Midtest NCE Mean	NCE Gain	N	t	p
A	II	44.6	45.6	1.1	21	.50	--
	III	35.8	39.4	3.6	32	1.21	--
	IV	31.0	36.9	5.8	30	3.22	.005
	Combined	36.3	40.1	3.8	83	2.65	.005
B	II	32.5	36.2	3.2	40	2.56	.01
	III	38.8	41.4	2.6	87	1.64	.05
	IV	32.0	34.3	2.2	41	1.43	.05
	Combined	35.6	38.4	2.8	168	2.87	.005
C	II	36.2	37.7	1.5	28	.57	--
	III	37.9	40.9	3.0	47	1.97	.05
	IV	38.2	41.2	3.0	53	1.93	.05
	Combined	37.6	40.3	2.7	128	2.63	.005
D	II	31.5	35.4	3.9	15	1.43	--
	III	32.5	37.4	5.0	52	2.39	.025
	IV	29.2	28.2	- .9	77	.66	--
	Combined	30.6	32.3	1.7	144	1.53	--
All	II	35.8	38.4	2.6	104	2.46	.01
	III	36.6	40.0	3.4	218	3.51	.001
	IV	32.4	34.2	1.8	201	2.15	.025
	Combined	34.8	37.5	2.6	523	4.75	.001

When the data are combined across cohorts within sites, Site A emerges with the largest pre-to-midtest gain. This finding is exactly the opposite of what one would expect based on what is known about implementation events at the various sites. The pre-to-posttest results, on the other hand, place the sites in approximately the predicted order.

The pre-to-posttest results show a marked improvement in performance with successive cohorts at Sites C and D and overall. Again, this finding is consistent with expectations based on implementation events. The large gain made by fourth-cohort interns at Site B, on the other hand, is counter-intuitive. One can only speculate that the disarray resulting from the director's departure did not affect the efficacy of instruction related to the development of reading skills.

Table 4
Treatment Group Pre-to-Posttest NCE Gains in Reading:
Estimates Derived from Norm-Referenced Analyses

Site	Cohort	Pretest NCE Mean	Posttest NCE Mean	NCE Gain	N	t	p
A	II	45.2	49.7	4.5	18	1.87	.05
	III	32.6	36.6	4.0	22	1.38	--
	IV	34.6	37.7	3.2	21	.87	--
	Combined	37.0	40.8	3.8	61	2.20	.025
B	II	34.0	37.2	3.2	15	1.23	--
	III	41.4	48.5	7.1	50	2.82	.005
	IV	32.4	40.8	8.4	32	4.68	.001
	Combined	37.3	44.2	6.9	97	4.67	.001
C	II	31.0	29.0	-2.0	9	.61	--
	III	34.0	39.6	5.6	21	1.64	--
	IV	39.6	48.3	8.7	34	4.98	.001
	Combined	36.5	42.7	6.2	64	3.92	.001
D	II	33.5	33.7	.2	6	.06	--
	III	34.8	42.3	7.5	33	2.54	.01
	IV	30.5	40.2	9.7	67	5.47	.001
	Combined	32.0	40.5	8.5	106	5.76	.001
All	II	37.6	39.9	2.3	48	1.61	--
	III	36.9	43.3	6.4	126	4.39	.001
	IV	33.5	41.8	8.3	154	7.79	.001
	Combined	35.4	42.1	6.7	328	8.51	.001

In general, the results of the norm-referenced reading analyses appear quite positive with the 328 students in the pre-to-posttest sample showing growth (on the average) from the 24th to the 35th percentile of the national distribution. This appearance of success, however, is somewhat lessened when one examines the norm-referenced gains made by control and comparison students. As shown in Tables 5 and 6, most of these groups also made statistically significant norm-referenced gains, some of which are actually larger than those made by the CIP participants.

Comparisons between the norm-referenced gain estimates for third-cohort treatment and control groups favor the treatment group at all four sites and overall at posttest time. For the fourth cohort, the treatment group out-performed the control group at two sites and overall. The midtest results are slightly less favorable. For the third cohort, treatment group gains are larger at three

Table 5
Control and Comparison Group Pre-to-Midtest NCE Gains in Reading: Estimates Derived from Norm-Referenced Analyses

Site	Cohort	Group	Pretest NCE Mean	Midtest NCE Mean	NCE Gain	N	t	p
A	III	Control	35.5	36.8	1.3	19	.34	--
	III	Reg. HS	46.7	49.0	2.3	55	.96	--
	III	Alt. HS	45.6	47.0	1.4	50	.59	--
	III	Dropout	47.6	46.4	- 1.2	19	.34	--
	IV	Control	34.4	35.6	1.2	27	.49	--
B	III	Control	35.2	38.6	3.4	25	1.39	--
	III	Reg. HS	32.7	36.8	4.2	51	3.00	.005
	III	Alt. HS	40.6	44.3	3.6	54	2.94	.005
	IV	Control	36.6	42.8	6.3	32	3.03	.005
C	III	Control	41.9	42.2	.3	30	.11	--
	III	Reg. HS	45.9	52.5	6.6	55	3.49	.005
	III	Alt. HS	57.1	56.5	- .6	39	.30	--
	IV	Control	42.5	41.0	- 1.6	10	.44	--
D	III	Control	32.4	34.2	1.8	15	.55	--
	IV	Control	33.3	35.9	2.6	54	1.48	--
All	III	Control	37.0	38.7	1.6	89	1.14	--
	III	Reg. HS	42.0	46.4	4.4	161	3.86	.001
	III	Alt. HS	46.9	48.6	1.7	143	1.52	--
	III	Dropout	47.6	46.4	- 1.2	19	.34	--
	IV	Control	35.2	38.1	2.9	123	2.59	.01

of the four sites and overall. The fourth-cohort control group, however, outgained the treatment group at two sites and overall. The midtest result at Site D is again difficult to accept at face value in view of what is known of implementation events at that site and the fact that the same group shows a very large gain at posttest time (9.7 NCEs).

In the case of the comparison groups, both the regular high school and the dropout groups outgained third-cohort CIP participants at posttest time. The regular high school group also outperformed the CIP group at midtest time.

Why the gains made by the regular high school group are so large is not clear. There is no reason to believe that these schools were doing an outstanding job teaching their students to read. A more plausible explanation is that some sort of selection took place--perhaps by the classroom teacher motivated to look good,

Table 6
Control and Comparison Group Pre-to-Posttest NCE Gains in Reading: Estimates Derived from Norm-Referenced Analyses

Site	Cohort	Group	Pretest NCE Mean	Posttest NCE Mean	NCE Gain	N	t	p
A	III	Control	32.5	34.6	2.1	16	.69	--
	III	Reg. HS	45.8	52.7	6.9	39	2.43	.025
	III	Alt. HS	46.7	41.7	-5.0	28	1.60	--
	III	Dropout	48.8	56.1	7.3	16	2.19	.025
	IV	Control	36.5	39.8	3.3	18	.83	--
B	III	Control	36.1	41.9	5.8	20	2.21	.025
	III	Reg. HS	32.9	41.9	9.0	42	5.46	.001
	III	Alt. HS	32.3	39.0	7.0	26	2.91	.005
	IV	Control	33.9	45.0	11.1	26	5.44	.001
C	III	Control	41.7	47.0	5.3	14	1.88	.05
	III	Reg. HS	48.3	55.3	7.0	51	3.27	.005
	III	Alt. HS	57.4	56.7	- .7	8	.09	--
	IV	Control	29.8	30.2	.4	12	.12	--
D	III	Control	31.0	34.6	3.6	16	1.36	--
	IV	Control	32.4	37.4	5.0	50	2.18	.025
All	III	Control	35.2	39.5	4.3	66	3.10	.005
	III	Reg. HS	42.7	50.3	7.6	132	5.93	.001
	III	Alt. HS	42.0	42.5	.5	62	.24	--
	III	Dropout	48.8	56.1	7.3	16	2.19	.025
	IV	Control	33.2	38.8	5.6	106	3.94	.001

perhaps by the students themselves--so that only the students who had shown improvement completed the mid- and posttests. Since there were 211 students in the regular high school group at pretest time and only 161 and 132 at mid- and posttest times respectively, this explanation is at least possible, if not particularly compelling.

The large gain made by the dropout group must be interpreted cautiously. With only 16 members in the group, the size of the gain could vary over a wide range. Although the differences were not tested, it is unlikely that the dropout group's gain is significantly different from that of the third-cohort treatment group either at Site A or overall.

Tables 7 and 8 present the results of the control group comparisons performed by means of covariance analysis (ANCOVA) on the reading test scores. The generally negative findings of these analyses are not inconsistent with those of the norm-referenced

Table 7
Treatment Group NCE Gains in Reading at Midtest Time:
Estimates Derived from Covariance Analyses

Site	Cohort	Group	Pretest Mean	Adj. Mid-test Mean	Gain	N	F	p
A	III	Treat.	35.8	39.3	2.4	32	.23	--
		Control	35.5	36.9		19		
	IV	Treat.	31.0	38.1	3.9	30	1.98	--
		Control	34.4	34.2		27		
	Comb.	Treat.	33.5	38.7	3.3	62	1.50	--
		Control	34.9	35.4		46		
B	III	Treat.	38.8	40.8	.1	87	.01	--
		Control	35.2	40.7		25		
	IV	Treat.	32.0	36.1	-4.3	41	2.80	--
		Control	36.6	40.5		32		
	Comb.	Treat.	36.6	39.0	-2.4	128	1.38	--
		Control	35.9	41.3		57		
C	III	Treat.	37.9	42.2	2.1	47	.65	--
		Control	41.9	40.1		30		
	IV	Treat.	38.2	41.8	4.1	53	1.11	--
		Control	42.5	37.7		10		
	Comb.	Treat.	38.0	42.0	2.7	100	1.64	--
		Control	42.0	39.4		40		
D	III	Treat.	32.5	37.4	3.1	52	.57	--
		Control	32.4	34.3		15		
	IV	Treat.	29.2	29.9	-3.7	77	2.74	--
		Control	33.3	33.6		54		
	Comb.	Treat.	30.5	32.8	-1.2	129	.37	--
		Control	33.1	34.0		69		
All	III	Treat.	36.2	40.2	2.0	218	1.43	--
		Control	37.0	38.2		89		
	IV	Treat.	32.4	35.1	-1.4	201	1.06	--
		Control	35.2	36.5		123		
	Comb.	Treat.	34.6	37.6	.1	419	.01	--
		Control	36.0	37.5		212		

Table 8
Treatment Group NCE Gains in Reading at Posttest Time:
Estimates Derived from Covariance Analyses

Site	Cohort	Group	Pretest Mean	Adj. Post-test Mean	Gain	N	F	p
A	III	Treat.	32.6	36.5	1.8	22	.18	--
		Control	32.5	34.7		16		
	IV	Treat.	34.6	38.3	- .7	21	.02	--
		Control	36.5	39.1		18		
	Comb.	Treat.	33.6	37.5	.6	43	.03	--
		Control	34.6	36.9		34		
B	III	Treat.	41.4	47.4	3.1	50	.50	--
		Control	36.2	44.3		20		
	IV	Treat.	32.4	41.4	-2.8	32	1.09	--
		Control	33.9	44.2		26		
	Comb.	Treat.	37.9	44.7	- .5	82	.04	--
		Control	34.9	45.2		46		
C	III	Treat.	34.0	42.5	- .3	21	.00	--
		Control	41.7	42.8		14		
	IV	Treat.	40.0	45.9	8.9	34	5.68	.01
		Control	29.8	37.0		12		
	Comb.	Treat.	37.4	44.6	4.5	55	2.43	--
		Control	36.2	40.1		26		
D	III	Treat.	34.8	41.3	4.7	33	1.08	--
		Control	31.0	36.6		16		
	IV	Treat.	30.5	40.9	4.5	67	2.52	--
		Control	32.4	36.4		50		
	Comb.	Treat.	31.9	41.0	4.4	100	3.36	.05
		Control	32.1	36.6		66		
All	III	Treat.	36.9	42.8	2.4	126	1.16	--
		Control	35.2	40.4		66		
	IV	Treat.	33.4	41.7	2.7	154	2.56	--
		Control	33.2	39.0		106		
	Comb.	Treat.	35.0	42.2	2.6	280	3.47	.05
		Control	34.0	39.6		172		

analyses. The gains are highly similar, in fact, to what would be obtained by subtracting the norm-referenced gains of the control groups from those of corresponding treatment groups.

No statistically significant ANCOVA gain estimates were found at midtest time. At posttest time only 2 of the 12 individual-site analyses produced statistically significant gain estimates, although the overall (across sites, across cohorts) estimate is also significant. This last finding, of course, is the most important as it verifies that treatment group students, on the average, outperformed control group students.

There is some reason to believe that the gain estimates derived from the ANCOVAs may be biased. Apart from the very high attrition rates in both treatment and control groups (which could have led to systematic differences between groups), the fact that all members of all control groups applied for, but were denied admission to the CIP may have had some effect on their motivation. Indeed, it seems likely that the so-called John Henry effect (Saretsky, 1972) may have been operating and may have artificially inflated the gains made by the various control groups.

Tables 9 and 10 present the results of the comparison group analyses derived through use of standardized gain procedures. As was the case with the covariance analyses, none of the gains was found to be statistically significant at midtest time. At posttest time, only one of the individual-site and one of the across-site estimates was significant. It is interesting to note that the one significant across-site gain involves the alternative high school comparison group, suggesting that the CIP is outperforming other programs serving similar youths.

Site B is, not surprisingly, an exception to this general trend. The entire alternative high school group at Site B was enrolled in a single program that provides intensive remedial reading instruction.

Overall, the comparison group analyses were marred by large initial differences between treatment and comparison groups. Although every effort was made to select low achievers, it is clear that this goal was only achieved at Site B (where, in fact, our efforts were somewhat too successful). At Sites A and C most comparison groups are only slightly below the national median (an NCE of 50) and one is substantially above it. With differences as large as these, any attempt at statistical equating requires assumptions of heroic proportions.

Tables 11 and 12 present the results of the matched-pairs analyses. While, in theory, these analyses might have provided the best insights relative to program impact, high attrition produced extremely small sample sizes. As a result, only Site C shows a significant gain at midtest time and only Site D at posttest time. The across-site gain at posttest time is also significant for the third cohort.

Table 9
Treatment Group NCE Gains in Reading at Midtest Time:
Estimates Derived from Standardized Gain Analyses, Third Cohort

Site	Group	Pretest Mean	Adj. Mid-test Mean	Gain	N	F	p
A	Treatment	35.8	47.2	2.7	32	.46	--
	Reg. HS	46.7	44.5		55		
	Treatment	35.8	46.6	4.2	32	1.11	--
	Alt. HS	45.6	42.4		50		
	Treatment	35.8	45.1	8.3	32	2.74	--
	Dropout	47.6	36.8		19		
B	Treatment	38.8	39.0	-2.0	87	.68	--
	Reg. HS	32.7	41.0		51		
	Treatment	38.8	42.2	- .9	87	.16	--
	Alt. HS	40.6	43.1		54		
C	Treatment	37.9	45.5	-3.1	47	1.48	--
	Reg. HS	45.9	48.6		55		
	Treatment	37.9	49.7	3.8	47	2.31	--
	Alt. HS	57.1	45.8		39		
All	Treatment	38.0	43.0	-1.1	166	.46	--
	Reg. HS	42.0	44.1		161		
	Treatment	38.0	45.3	1.9	166	1.42	--
	Alt. HS	46.9	43.4		143		
	Treatment	35.8	45.1	8.3	32	2.74	--
	Dropout	47.6	36.8		19		

Table 10
Treatment Group NCE Gains in Reading at Posttest Time:
Estimates Derived from Standardized Gain Analyses, Third Cohort

Site	Group	Pretest Mean	Adj. Post-test Mean	Gain	N	F	p
A	Treatment	32.6	46.7	- .3	22	.00	--
	Reg. HS	45.8	47.0		39		
	Treatment	32.6	47.1	13.7	22	9.10	.005
	Alt. HS	46.6	33.4		28		
	Treatment	32.6	45.5	1.8	22	.16	--
	Dropout	48.8	43.7		16		
B	Treatment	41.4	44.0	-3.3	50	.97	--
	Reg. HS	32.9	47.3		42		
	Treatment	41.4	44.9	- .9	50	.05	--
	Alt. HS	32.3	45.8		26		
C	Treatment	34.0	50.2	- .7	21	.03	--
	Reg. HS	48.3	50.9		51		
	Treatment	34.0	48.0	13.1	21	2.87	--
	Alt. HS	57.4	34.9		8		
All	Treatment	37.6	47.0	- .9	93	.19	--
	Reg. HS	42.7	47.9		132		
	Treatment	37.6	45.8	6.5	93	5.23	.01
	Alt. HS	42.0	39.3		62		
	Treatment	32.6	45.5	1.8	22	.16	--
	Dropout	48.8	43.7		16		

Table 11
Treatment Group NCE Gains in Reading at Midtest Time:
Estimates Derived from Matched Pairs Analyses

Site	Cohort	Mean Mid-test NCE Treatment	Mean Mid-test NCE Control	NCE Gain	N	t	p
A	III	33.9	26.4	7.5	7	1.05	--
	IV	39.6	34.6	4.9	7	1.07	--
	Combined	36.7	30.5	6.2	14	1.53	--
B	III	35.9	37.2	-1.3	23	.39	--
	IV	43.3	48.1	-4.8	13	.86	--
	Combined	38.6	41.2	-2.6	36	.88	--
C	III	42.6	36.4	6.2	17	1.81	.05
	IV	48.2	41.6	6.7	9	1.07	--
	Combined	44.5	38.2	6.3	26	2.10	.025
D	III	38.1	34.1	4.0	13	1.13	--
	IV	29.9	32.7	-2.8	23	.78	--
	Combined	32.9	33.2	- .3	36	.12	--
All Sites	III	38.0	35.0	3.0	60	1.51	--
	IV	37.7	38.4	- .6	52	.25	--
	Combined	37.9	36.6	1.3	112	.83	--

By far the most positive results with respect to reading achievement are observed in the norm-referenced analyses. Surprisingly, however, the norm-referenced gain estimates for most of the control and comparison groups are also positive rather than zero as might have been expected (at least for the regular high school comparison groups). If these control and comparison group gains are "real," then the norm-referenced analyses produced the most valid gain estimates. The possibility must be acknowledged, however, that these gains are no more than artifacts of the norm-referenced procedures employed in the evaluation.

Table 12
Treatment Group NCE Gains in Reading at Posttest Time:
Estimates Derived from Matched Pairs Analyses

Site	Cohort	Mean Post-test NCE Treatment	Mean Post-test NCE Control	NCE Gain	N	t	p
A	III	38.7	29.9	8.8	2	.58	--
	IV	39.9	32.6	7.4	5	.98	--
	Combined	39.6	31.8	7.8	7	1.27	--
B	III	41.6	42.4	- .8	14	.16	--
	IV	41.3	52.9	-11.6	9	1.90	--
	Combined	41.5	46.5	- 5.0	23	1.21	--
C	III	56.0	56.2	- .2	4	.02	--
	IV	39.3	27.8	11.5	5	1.16	--
	Combined	46.7	40.4	6.3	9	.98	--
D	III	55.7	43.3	12.4	7	1.85	--
	IV	45.6	39.9	5.6	17	1.18	--
	Combined	48.5	40.9	7.6	24	1.96	.05
All Sites	III	47.2	43.8	3.4	27	3.39	.005
	IV	42.9	40.5	2.4	36	.70	--
	Combined	44.7	41.9	2.8	63	1.14	--

As mentioned earlier, it was necessary to implement the norm-referenced model in a somewhat unorthodox manner. Two specific deviations from standard implementation procedures could have introduced some distortions. The first deviation was the extrapolation and interpolation of normative data to accommodate the flexible testing schedule imposed on the study by various practical considerations. The second deviation was the assignment of students to grade-level norms on the basis of age rather than their actual grade placement. Either of these procedural variations could have introduced bias into the analyses. The authors, however, are unable to generate a plausible explanation as to why the bias should have been consistently positive regardless of testing times or type of group. We are inclined instead to favor the hypothesis that the norm-referenced gain estimates are accurate and that the gains apparently made by the control and comparison groups resulted from some combination of the John Henry effect and a selection bias. In any case, it should be remembered that, overall, the treatment group significantly outperformed the control group and the alternative high school group.

Math

Tables 13 and 14 summarize the pre-to-midtest and pre-to-posttest norm-referenced analyses of mathematics test scores. As can be seen, nearly half of the gain estimates at midtest time are statistically significant and a majority are significant at posttest time. Combined across sites and cohorts, the mean pre-to-midtest gain is 2.2 NCEs and the pre-to-posttest gain is 4.3 NCEs. The latter gain is somewhat smaller than that observed for reading achievement, a finding that is readily explainable in terms of the difficulty all of the sites experienced in hiring and retaining qualified math instructors.

Table 13
Treatment Group Pre-to-Midtest NCE Gains in Math:
Estimates Derived from Norm-Referenced Analyses

Site	Cohort	Pretest NCE Mean	Midtest NCE Mean	NCE Gain	N	t	p
A	II	31.2	32.5	1.3	21	.36	--
	III	19.2	26.5	7.3	32	2.14	.025
	IV	25.2	27.4	2.3	30	.79	--
	Combined	24.4	28.4	4.0	83	2.01	.025
B	II	23.4	24.9	1.5	40	1.49	--
	III	27.3	30.0	2.7	87	1.74	.05
	IV	24.9	27.2	2.2	41	1.58	--
	Combined	25.8	28.1	2.3	168	2.45	.01
C	II	31.6	30.7	- .9	28	.32	--
	III	31.0	34.7	3.6	46	2.59	.01
	IV	31.2	31.8	.7	53	.45	--
	Combined	31.2	32.6	1.4	127	1.36	--
D	II	26.0	30.2	4.2	14	1.41	--
	III	23.7	26.9	3.2	48	2.37	.025
	IV	23.6	23.8	.3	77	.19	--
	Combined	23.9	25.5	1.7	139	1.77	.05
All	II	27.6	28.8	1.2	103	.91	--
	III	26.1	29.8	3.7	213	4.03	.001
	IV	26.1	27.2	1.1	201	1.29	--
	Combined	26.4	28.6	2.2	517	6.91	.001

Table 14
Treatment Group Pre-to-Posttest NCE Gains in Math:
Estimates Derived from the Norm-Referenced Analyses

Site	Cohort	Pretest NCE Mean	Posttest NCE Mean	NCE Gain	N	t	p
A	II	30.6	36.8	6.3	18	1.75	.05
	III	16.1	30.1	14.0	22	3.28	.005
	IV	23.6	22.3	- 1.3	21	.61	--
	Combined	23.0	29.4	6.4	61	3.0	.005
B	II	20.8	24.7	3.9	15	1.29	--
	III	27.1	36.0	8.9	50	4.42	.001
	IV	25.7	25.4	- .3	32	.11	--
	Combined	25.7	30.8	5.1	97	3.66	.001
C	II	27.5	24.2	- 3.2	9	.97	--
	III	28.9	29.1	.2	21	.07	--
	IV	32.2	38.5	6.3	34	2.92	.005
	Combined	30.5	33.4	3.0	64	1.85	.05
D	II	20.9	29.7	8.8	6	1.68	--
	III	25.5	30.5	5.0	32	2.45	.025
	IV	24.0	25.8	1.7	67	1.00	--
	Combined	24.3	27.4	3.1	65	.26	--
All	II	25.8	29.8	4.0	48	2.14	.025
	III	25.1	32.4	7.3	125	5.44	.001
	IV	26.1	28.0	1.9	154	1.81	.05
	Combined	25.7	30.0	4.3	327	5.52	.001

The pre-to-midtest results, when combined across sites, show that the smallest gain was made by fourth-cohort students. Again it seems likely that this finding is best explained by the short pre-to-midtest interval for this cohort. The largest gain was made by third-cohort students--a not surprising outcome in view of the fact that there were fewer implementation problems during the time period in question than was the case during the tenure of either the second- or fourth-cohorts. What is surprising is that the largest gain was made at Site A. However, despite other problems at that site, it did have an excellent math teacher.

The authors were initially somewhat concerned about the quite low mean pretest score for the third-cohort group at Site A--especially since the corresponding score for the control group is 11.5 NCEs higher. Initially we thought that there might have been a few invalid scores that would not only account for the low pretest mean but also for the large gains both from pre- to midtest and from

pre- to posttest. Examination of the raw data, however, revealed no such problem. The difference between the pretest scores of the treatment and control groups appears to be the result of high-scoring members of the treatment group failing to enroll in the CIP or dropping out before midtest time. (The pretest means of the two groups prior to attrition were 25.8 and 26.1 NCEs respectively.) The reality of the pre- to midtest gain, furthermore, is attested to by the continued growth from mid- to posttest which could not result from invalid pretest scores.

When the data are combined across cohorts within each site, Site A emerges with the largest pre-to-midtest and pre-to-posttest gain. In both cases, the third cohort is primarily responsible. The outstanding math instructor was not hired until some time after the second cohort enrolled and left before the fourth cohort entered the program.

The trend toward improvement over time that was observed in reading at both Sites C and D is seen only at Site C in math. At Site D the gain made by the fourth cohort was less than that made by the third. This reversal is attributed to the departure of the site's excellent science teacher who also often taught math classes. His departure more than offset the general improvement in climate that was reported earlier.

The pre-to-posttest results show much the same pattern that was observed at midtest time. However, when summarized across sites, the gains made by all three cohorts are statistically significant. The 5 NCE gain made by second-cohort students at Site A adds further credibility to the effectiveness of the math instructor at that site. She joined the program just before the second cohort was midtested. Similarly, the 6.2 NCE gain made by third-cohort interns at Site B between mid- and posttests can be attributed to the fact that a well qualified and talented math instructor was finally hired at that site. Unfortunately, he left again after only six months.

Overall, the norm-referenced results are encouraging. They also suggest that larger gains would have occurred had math teaching positions been vacant less often. In any case, the 327 students who had both pre- and posttests moved from a national percentile rank of 12.4 to 17.1. It is perhaps noteworthy that the math achievement of CIP students is substantially below the level in reading.

Tables 15 and 16 present summaries of the norm-referenced analyses performed on control and comparison group math achievement data. Only a few of these gain estimates are statistically significant and most of them are smaller than those made by the corresponding treatment groups. Summarized across sites, none of the control or comparison group gains at midtest time exceed those made by the corresponding treatment group. The same situation prevails at posttest time, with the single exception that the fourth-cohort control group outgained the treatment group by .2 NCEs.

Table 15
Control and Comparison Group Pre-to-Midtest NCE Gains in Math: Estimates Derived from Norm-Referenced Analyses

Site	Cohort	Group	Pretest NCE Mean	Midtest NCE Mean	NCE Gain	N	t	p
A	III	Control	30.7	28.7	-2.0	19	.59	--
	III	Reg. HS	41.3	48.1	6.8	54	2.33	.025
	III	Alt. HS	37.6	40.0	2.3	50	.89	--
	III	Dropout	40.4	42.5	2.1	19	.62	--
	IV	Control	25.4	21.9	-3.5	27	1.46	--
B	III	Control	28.9	34.7	5.8	25	1.61	--
	III	Reg. HS	35.0	36.8	1.8	51	1.92	.025
	III	Alt. HS	38.2	37.1	-1.0	53	.61	--
	IV	Control	28.9	27.7	-1.2	32	.62	--
C	III	Control	26.6	24.8	-1.9	30	.82	--
	III	Reg. HS	41.8	43.9	2.1	55	1.08	--
	III	Alt. HS	48.5	51.2	2.7	39	1.34	--
	IV	Control	32.0	34.0	2.1	10	.81	--
D	III	Control	29.1	32.4	3.3	14	.71	--
	IV	Control	26.4	27.3	.9	54	.42	--
All	III	Control	28.5	29.6	1.1	88	.67	--
	III	Reg. HS	39.5	43.1	3.6	160	2.92	.005
	III	Alt. HS	40.8	42.0	1.2	142	.93	--
	III	Dropout	40.4	42.5	2.1	19	.62	--
	IV	Control	27.3	26.7	- .5	123	.43	--

At posttest time, the smallest gain (-3.7 NCEs) was registered by the alternative high school comparison group, suggesting a real superiority of the CIP compared to other like programs. This difference is most marked at Site A where the CIP is only one of several alternative programs in the school district.

The treatment-control analyses performed using covariance analysis are summarized in Tables 17 and 18. Only three of the gain estimates are statistically significant at midtest time and none is significant at posttest time. The larger treatment group gains made by third-cohort students at Site B and by fourth-cohort students at Site C are largely offset by the sizeable gains registered by the corresponding control groups. Again, selection biases and John Henry effects may have been operative.

Table 16
Control and Comparison Group Pre-to-Posttest NCE Gains in Math: Estimates Derived from Norm-Referenced Analyses

Site	Cohort	Group	Pretest NCE Mean	Posttest NCE Mean	NCE Gain	N	t	p
A	III	Control	29.4	31.6	2.2	16	2.19	.025
	III	Reg. HS	42.2	42.8	.6	39	.29	--
	III	Alt. HS	42.6	33.1	-9.5	28	2.39	--
	III	Dropout	39.6	41.3	1.7	16	.37	--
	IV	Control	28.2	32.8	4.7	18	.91	--
B	III	Control	26.4	30.5	4.1	20	2.95	.005
	III	Reg. HS	37.5	42.7	5.2	42	3.29	.005
	III	Alt. HS	35.5	36.7	1.2	26	.42	--
	IV	Control	29.0	31.8	2.8	26	1.15	--
C	III	Control	33.2	37.5	4.3	13	1.28	--
	III	Reg. HS	42.6	43.6	1.0	51	.60	--
	III	Alt. HS	46.2	47.2	1.0	8	.31	--
	IV	Control	24.2	29.8	5.6	11	1.44	--
D	III	Control	25.3	27.1	1.8	15	.41	--
	IV	Control	26.0	26.0	--	50	--	--
All	III	Control	28.3	31.5	3.2	65	1.93	.05
	III	Reg. HS	40.9	43.1	2.2	132	2.12	.025
	III	Alt. HS	40.1	36.4	-3.7	62	1.60	--
	III	Dropout	39.6	41.3	1.7	16	.37	--
	IV	Control	26.9	29.0	2.1	105	1.47	--

At Site A, where the third-cohort norm-referenced achievement gain is very large, it is somewhat surprising to find that the covariance analyses shows a substantially smaller and statistically non-significant gain--especially since the control group's (norm-referenced) gain is comparatively small (2.2 NCEs). In fact, the apparent inconsistency stems from the large difference between the pretest scores of the two groups. It is clear that the two groups were equivalent initially but experienced systematically different types of attrition. Students who remained in the treatment group at mid- and posttest times were clearly different from those who remained in the control group at the same times.

Real differences between groups result in a systematic under-correction of posttest scores when traditional ANCOVA procedures are used (Campbell & Boruch, 1975). In this particular case at least, it appeared that a more valid gain estimate would be obtained using standardized gain analysis. When this was done, a pre-to-midtest

Table 17
Treatment Group NCE Gains in Math at Midtest Time:
Estimates Derived from Covariance Analyses

Site	Cohort	Group	Pretest Mean	Adj. Mid-test Mean	Gain	N	F	p
A	III	Treat.	19.2	29.5	5.8	32	1.37	--
		Control	30.7	23.7		19		
	IV	Treat.	25.2	27.5	5.7	30	2.40	--
		Control	25.4	21.8		27		
	Comb.	Treat.	22.1	28.6	6.2	62	4.35	.025
		Control	27.6	22.4		46		
B	III	Treat.	27.3	30.3	-3.6	87	1.23	--
		Control	28.9	33.9		25		
	IV	Treat.	24.9	28.5	2.6	41	1.32	--
		Control	28.9	26.0		32		
	Comb.	Treat.	26.6	29.6	--	128	--	--
		Control	28.9	29.6		57		
C	III	Treat.	31.0	33.2	6.2	46	6.08	.001
		Control	26.6	27.0		30		
	IV	Treat.	31.2	31.9	-1.6	53	.24	--
		Control	32.0	33.5		10		
	Comb.	Treat.	31.1	32.4	3.5	99	3.39	.05
		Control	28.0	28.9		40		
D	III	Treat.	23.7	28.0	- .9	48	.07	--
		Control	29.1	28.9		14		
	IV	Treat.	23.6	24.7	-1.3	77	.29	--
		Control	26.4	26.0		54		
	Comb.	Treat.	23.6	26.0	- .7	125	.12	--
		Control	26.9	26.6		68		
All	III	Treat.	26.1	30.3	2.0	213	1.34	--
		Control	28.5	28.4		88		
	IV	Treat.	26.1	27.5	1.3	201	.98	--
		Control	27.3	26.2		123		
	Comb.	Treat.	26.1	29.0	1.9	414	3.03	.05
		Control	27.8	27.1		211		

Table 18
Treatment Group NCE Gains in Math at Posttest Time:
Estimates Derived from Covariance Analyses

Site	Cohort	Group	Pretest Mean	Adj. Post-test Mean	Gain	N	F	p
A	III	Treat.	16.1	33.6	6.8	22	1.28	--
		Control	29.4	26.8		16		
	IV	Treat.	23.6	23.6	-7.6	21	2.32	--
		Control	28.2	31.2		18		
	Comb.	Treat.	19.8	28.6	- .7	43	.03	--
		Control	28.7	29.3		34		
B	III	Treat.	27.1	35.8	4.9	50	2.9	--
		Control	26.4	30.9		20		
	IV	Treat.	25.7	26.6	-3.8	32	1.45	--
		Control	29.0	30.4		26		
	Comb.	Treat.	26.6	32.3	1.7	82	.54	--
		Control	27.9	30.6		46		
C	III	Treat.	28.9	30.7	-4.6	21	1.02	--
		Control	33.2	35.3		14		
	IV	Treat.	32.2	37.3	3.5	34	.77	--
		Control	24.3	33.8		11		
	Comb.	Treat.	30.9	34.5	- .5	55	.03	--
		Control	29.3	35.0		25		
D	III	Treat.	25.3	30.5	3.3	32	.68	--
		Control	25.3	27.2		15		
	IV	Treat.	24.0	26.4	1.2	67	.25	--
		Control	26.0	25.2		50		
	Comb.	Treat.	24.5	27.7	2.0	99	.93	--
		Control	25.8	25.7		65		
All	III	Treat.	25.1	33.2	3.3	125	2.46	--
		Control	28.3	29.9		65		
	IV	Treat.	26.1	28.3	- .4	154	.05	--
		Control	26.9	28.7		105		
	Comb.	Treat.	25.6	30.5	1.4	279	1.08	--
		Control	27.5	29.1		170		

gain estimate of 9.6 NCEs was obtained. The pre-to-posttest gain was 12.7 NCEs. Both estimates are statistically significant at the .05 level.

Standardized gain analysis was also applied to the across-site comparison between third-cohort treatment and control groups. This approach raised the gain estimate from 3.3 to 4.6 NCEs and the latter value is significant at the .025 level. When third and fourth cohorts were combined, the standardized gain estimate rose to 2 NCEs but remained statistically non-significant.

Tables 19 and 20 present the results of the standardized gain analyses performed on treatment and comparison group data. At mid-test time, only one of the ten gain estimates is statistically significant. At posttest time, on the other hand, only two of ten fail to attain statistical significance. It should be noted, however, that the credibility of these highly positive results is substantially diminished by the very large pretest differences between groups. Although there is no reason to believe that the analysis methodology introduced biases in either direction, it is simply not very informative to make comparisons between groups that have so little in common. While the fact that the results are positive does provide some further evidence supporting the success of the CIP, the gain estimates themselves appear badly inflated--particularly at Site A and across sites.

The matched pairs analyses, presented in Tables 21 and 22, are equally uninformative. Only 1 of 30 is significant at the .025 level--an event not unlikely to occur by chance.

Table 19
Treatment Group NCE Gains in Math at Midtest Time:
Estimates Derived from Standardized Gain Analyses, Third Cohort

Site	Group	Pretest Mean	Adj. Mid-test Mean	Gain	N	F	p
A	Treatment	19.2	42.1	3.2	32	.44	--
	Reg. HS	41.3	38.9		54		
	Treatment	19.2	38.7	6.6	32	2.18	--
	Alt. HS	37.6	32.1		50		
	Treatment	19.2	34.2	4.8	32	.87	--
	Dropout	40.4	29.4		19		
B	Treatment	27.3	33.1	1.6	87	.53	--
	Reg. HS	35.0	31.5		51		
	Treatment	27.3	34.2	4.0	87	2.74	--
	Alt. HS	38.2	30.2		53		
C	Treatment	31.0	41.2	2.8	46	1.21	--
	Reg. HS	41.8	38.4		55		
	Treatment	31.0	43.7	3.1	46	1.54	--
	Alt. HS	48.5	40.6		39		
All	Treatment	26.8	37.6	1.7	165	.92	--
	Reg. HS	39.5	35.9		160		
	Treatment	26.8	37.6	3.7	165	4.50	.025
	Alt. HS	40.8	33.9		142		
	Treatment	19.2	34.2	4.8	32	.87	--
	Dropout	40.4	29.4		19		

Table 20
Treatment Group NCE Gains in Math at Posttest Time:
Estimates Derived from Standardized Gain Analyses, Third Cohort

Site	Group	Pretest Mean	Adj. Post-test Mean	Gain	N	F	p
A	Treatment	16.1	47.2	14.1	22	10.50	.005
	Reg. HS	42.2	33.1		39		
	Treatment	16.1	47.6	28.3	22	20.16	.001
	Alt. HS	42.6	19.3		28		
	Treatment	16.1	40.8	14.1	22	4.44	.025
	Dropout	39.6	26.7		16		
B	Treatment	27.1	41.6	5.6	50	4.10	.025
	Reg. HS	37.5	36.0		42		
	Treatment	27.1	39.2	8.7	50	5.73	.025
	Alt. HS	35.5	30.5		26		
C	Treatment	28.9	38.8	- .8	21	.06	--
	Reg. HS	42.6	39.6		51		
	Treatment	28.9	34.8	2.5	21	.23	--
	Alt. HS	46.2	32.3		8		
All	Treatment	24.9	42.9	6.8	93	12.42	.001
	Reg. HS	40.9	36.1		132		
	Treatment	24.9	40.0	13.2	93	22.83	.001
	Alt. HS	40.1	26.0		62		
	Treatment	16.1	40.8	14.1	22	4.44	.025
	Dropout	39.6	26.7		16		

Table 21
Treatment Group NCE Gains in Math at Midtest Time:
Estimates Derived from Matched Pairs Analyses

Site	Cohort	Mean Mid-test NCE Treatment	Mean Mid-test NCE Control	NCE Gain	N	t	p
A	III	17.2	15.7	1.5	7	.18	--
	IV	27.3	26.2	1.1	7	.13	--
	Combined	22.2	20.9	1.3	14	.23	--
B	III	32.0	33.9	-1.9	22	.40	--
	IV	31.9	27.5	4.4	17	1.30	--
	Combined	31.9	31.1	.8	39	.27	--
C	III	32.1	25.2	6.9	13	2.24	.025
	IV	23.1	29.8	-6.7	5	2.05	--
	Combined	29.6	26.4	3.2	18	1.13	--
D	III	34.2	38.7	-4.5	5	.48	--
	IV	23.2	27.5	-4.3	19	1.10	--
	Combined	25.5	29.8	-4.3	24	1.22	--
All Sites	III	30.0	29.3	.8	47	.27	--
	IV	26.9	27.5	- .7	48	.29	--
	Combined	28.4	28.4	.0	95	.02	--

Table 22
Treatment Group NCE Gains in Math at Posttest Time:
Estimates Derived from Matched Pairs Analyses

Site	Cohort	Mean Post-test NCE Treatment	Mean Post-test NCE Control	NCE Gain	N	t	p
A	III	23.4	15.6	7.8	3	.65	--
	IV	26.7	43.5	-16.8	3	1.57	--
	Combined	25.1	29.6	4.5	6	.50	--
B	III	27.8	33.3	- 5.5	11	1.24	--
	IV	30.6	35.7	- 5.1	10	.66	--
	Combined	29.2	34.4	- 5.3	21	1.26	--
C	III	49.2	42.6	6.6	2	1.56	--
	IV	33.4	26.1	7.3	5	.76	--
	Combined	37.9	30.8	7.1	7	1.06	--
D	III	24.5	32.4	- 7.9	6	1.05	--
	IV	19.1	18.3	.8	12	.18	--
	Combined	20.9	23.0	- 2.1	18	.58	--
All Sites	III	28.3	31.5	- 3.2	22	.94	--
	IV	26.1	27.9	- 1.8	30	.52	--
	Combined	27.0	29.4	- 2.4	52	.97	--

Career Development Inventory

Table 23 presents the pre-to-midtest raw score gains made by second-cohort students. Large gains were achieved on the CDI Planning scale by students at all sites at both mid- and posttest time. Except in two cases where the sample sizes were very small, these gains are all statistically significant. Much the same picture can be observed with the CDI Resources scale although the gains are somewhat smaller and four of them are non-significant.

Table 23
Treatment Group Pre-to-Midtest Raw Score Gains:
Career Development Inventory, Second Cohort

	Pretest Mean	Midtest Mean	Gain	N	t	p
Site A						
Planning	99.7	115.6	16.0	22	1.98	.05
Resources	76.3	82.6	6.3	22	1.21	--
Information	11.4	13.1	1.7	22	1.53	--
Site B						
Planning	100.0	121.0	21.0	37	6.01	.001
Resources	82.0	88.0	6.0	37	1.96	.05
Information	11.7	14.2	2.5	37	3.06	.005
Site C						
Planning	100.4	114.1	13.7	28	2.05	.025
Resources	82.8	90.9	8.1	28	1.30	--
Information	12.0	13.3	1.3	28	1.81	.05
Site D						
Planning	109.6	129.5	19.9	15	5.88	.001
Resources	86.1	95.9	9.8	15	3.02	.005
Information	14.5	15.5	1.0	15	1.06	--
All Sites						
Planning	101.6	119.5	17.9	102	6.25	.001
Resources	81.7	88.9	7.2	102	3.07	.005
Information	12.1	13.9	1.8	102	3.97	.001

Table 24
Treatment Group Pre-to-Posttest Raw Score Gains:
Career Development Inventory, Second Cohort

	Pretest Mean	Posttest Mean	Gain	N	t	p
Site A						
Planning	102.6	127.7	25.1	18	4.24	.001
Resources	75.4	84.1	8.6	17	1.32	--
Information	10.9	14.3	3.4	18	2.70	.01
Site B						
Planning	98.7	125.0	26.3	15	4.56	.001
Resources	76.4	87.7	11.3	15	2.54	.025
Information	12.9	14.8	1.9	15	1.61	--
Site C						
Planning	101.2	115.3	14.1	9	1.35	--
Resources	76.7	92.6	15.9	9	4.50	.005
Information	11.7	10.4	- 1.2	9	.87	--
Site D						
Planning	99.0	128.0	29.0	1	--	--
Resources	72.0	124.0	52.0	1	--	--
Information	16.2	19.0	2.8	6	2.10	.05
All Sites						
Planning	100.9	124.2	23.3	43	6 08	.001
Resources	75.9	88.0	12.1	43	3.63	.005
Information	12.3	14.3	2.0	48	2.85	.005

The CDI Information scale shows significant pre-to-midtest gains at Sites B and C and significant pre-to-posttest gains at Sites A and D. Across sites the gains on this scale are significant at both mid- and posttest times.

In the absence of both normative data and control groups, no other analyses of these data appear worth undertaking. It is important to note, however, that the analyses which are reported may be misleading. There would almost certainly be some growth over time without the CIP treatment. This growth, unfortunately, is inextricably confounded with whatever gains resulted from the treatment.

Tables 25 and 26 present gain estimates and related statistics derived from covariance analyses of treatment and control group scores on the CDI Planning scale. Table 25 summarizes the pre-to-midtest findings while Table 26 encompasses the pre-to-posttest

Table 25
Treatment Group Raw Score Gains
on the CDI Planning Scale at Midtest Time:
Estimates Derived from Covariance Analyses

Site	Cohort	Group	Pretest Mean	Adj. Mid-test Mean	Gain	N	F	p
A	III	Treat.	90.0	115.2	9.6	26	1.26	--
		Control	114.8	105.6		13		
	IV	Treat.	103.3	105.6	5.7	29	.45	--
		Control	100.6	100.0		26		
	Comb.	Treat.	97.0	110.4	9.0	55	2.25	--
		Control	105.4	101.4		39		
B	III	Treat.	105.2	120.7	9.6	80	4.67	.025
		Control	99.6	111.1		22		
	IV	Treat.	93.0	111.6	7.7	40	2.81	.05
		Control	95.5	103.9		30		
	Comb.	Treat.	101.1	117.3	9.5	120	9.4	.015
		Control	97.3	107.8		52		
C	III	Treat.	95.6	110.7	3.9	47	.56	--
		Control	103.7	106.8		29		
	IV	Treat.	103.7	102.3	-6.5	52	.36	--
		Control	100.2	108.8		10		
	Comb.	Treat.	99.9	106.2	-1.5	99	.09	--
		Control	102.8	107.7		39		
D	III	Treat.	106.3	121.4	8.0	47	1.91	--
		Control	107.9	113.4		15		
	IV	Treat.	110.1	118.2	7.8	72	4.23	.025
		Control	111.1	110.5		50		
	Comb.	Treat.	108.6	119.6	8.6	119	7.72	.005
		Control	110.4	111.0		65		
All	III	Treat.	101.2	117.8	8.6	200	10.25	.001
		Control	105.2	109.2		79		
	IV	Treat.	103.8	110.7	4.4	193	2.29	--
		Control	103.8	106.2		116		
	Comb.	Treat.	102.5	114.3	6.9	393	11.72	.001
		Control	104.4	107.4		195		

Table 26
Treatment Group Raw Score Gains
on the CDI Planning Scale at Posttest Time:
Estimates Derived from Covariance Analyses

Site	Cohort	Group	Pretest Mean	Adj. Post-test Mean	Gain	N	F	p
A	III	Treat.	91.7	122.1	9.8	18	.91	--
		Control	106.1	112.3		16		
	IV	Treat.	98.9	113.5	- .7	20	.01	--
		Control	108.3	114.2		16		
	Comb.	Treat.	95.4	117.5	4.1	38	.37	--
		Control	107.2	113.4		32		
B	III	Treat.	102.9	129.1	23.5	45	12.13	.001
		Control	99.1	105.6		17		
	IV	Treat.	94.3	112.4	3.8	32	.73	--
		Control	93.8	108.6		25		
	Comb.	Treat.	99.3	121.8	13.7	77	10.96	.005
		Control	95.3	108.1		42		
C	III	Treat.	91.4	117.1	4.9	18	.23	--
		Control	104.5	112.2		13		
	IV	Treat.	106.3	105.7	11.6	34	1.50	--
		Control	88.0	94.1		11		
	Comb.	Treat.	101.2	109.9	6.5	52	.93	--
		Control	96.9	103.4		24		
D	III	Treat.	105.5	125.8	17.9	30	7.67	.005
		Control	106.1	107.9		14		
	IV	Treat.	104.7	120.1	6.3	61	1.44	--
		Control	109.1	114.4		45		
	Comb.	Treat.	105.0	122.4	9.5	91	5.18	.05
		Control	108.4	112.9		59		
All	III	Treat.	99.9	125.2	15.9	111	16.71	.001
		Control	103.8	109.3		60		
	IV	Treat.	102.0	114.4	3.8	147	1.40	--
		Control	102.6	110.6		97		
	Comb.	Treat.	101.1	119.1	9.2	258	13.06	.001
		Control	103.1	109.9		157		

results. At midtest time, 5 of 12 individual-site gain estimates are statistically significant; at posttest time, only 4. Across sites, the third-cohort and the combined third-and-fourth-cohort gain estimates are statistically significant.

In several cases (e.g., Site A, third cohort, mid- and posttest; Site C, fourth cohort, posttest), there are large pretest differences between treatment and control groups which suggest the possibility that ANCOVA may be an inappropriate analytic approach. When standardized gain analyses were undertaken, three of the gains that were nonsignificant in the ANCOVAs attained statistical significance. These gains are as follows: (a) Site A third-cohort midtest--20.7 ($F = 5.54$, $p < .025$), (b) Site A combined third-and-fourth-cohort midtest--14.6 ($F = 4.44$, $p < .025$), and (c) Site A third-cohort posttest--22.95 ($F = 3.05$, $p < .05$). None of the other non significant ANCOVA estimates attained significance when standardized gain analyses were undertaken, but all of the significant ANCOVA estimates remained so, lending increased credibility to those findings.

There do not appear to be any meaningful differences among sites. On the other hand, the difference between third and fourth cohorts does appear meaningful. Except at Site C (where the fourth-cohort ANCOVA gain estimate at posttime is distorted by the very low pretest score of the control group), the same pattern is evident that is seen in the across-site comparisons. The lower fourth-cohort gain is attributable to the fact that student-counselor interactions were less frequent during the extension portion of the demonstration period than during the first two years. This reduction, in turn, is due to a number of career counselors leaving the program and others becoming overloaded with the paperwork created by large fourth-cohort enrollments, the inclusion of additional school districts in the recruitment/catchment area, and related problems.

Tables 27 and 28 summarize the standardized gain analyses performed on treatment and comparison group CDI Planning scores at midtest and posttest times respectively. Most of the gain estimates are both large and statistically significant both at midtest and posttest time. No clear patterns emerge with respect either to sites or comparison groups. There does, however, appear to be some continued growth from mid- to posttest.

Tables 29 and 30 summarize the ANCOVA results for the CDI Resources scale at mid- and posttest times respectively. At midtest time only one individual-site and none of the across-site gain estimates is statistically significant. As was the case with the CDI Planning scale, however, there are substantial pretest differences between treatment and control groups in a number of instances, suggesting that standardized gain analyses might yield more valid gain estimates than covariance analyses. When such analyses were carried out, the third-cohort gain estimate at Site C increased to 12.6 and became statistically significant ($F = 6.74$, $p <$

Table 27
Treatment Group Raw Score Gains
on the CDI Planning Scale at Midtest Time:
Estimates Derived from Standardized Gain Analyses, Third Cohort

Site	Group	Pretest Mean	Adj. Mid-test Mean	Gain	N	F	p
A	Treatment	90.0	123.5	18.3	26	7.69	.005
	Reg. HS	106.4	105.2		56		
	Treatment	90.0	130.0	34.3	26	20.37	.001
	Alt. HS.	112.1	95.7		49		
	Treatment	90.0	117.2	19.2	26	5.92	.025
	Dropout	106.8	98.0		18		
B	Treatment	105.2	120.6	10.3	80	7.69	.005
	Reg. HS	103.2	110.3		53		
	Treatment	105.2	122.0	14.5	80	11.21	.001
	Alt. HS	106.3	107.5		52		
C	Treatment	95.6	86.2	2.3	47	.32	--
	Reg. HS	57.6	83.9		55		
	Treatment	95.6	88.3	11.5	47	4.87	.025
	Alt. HS	52.4	76.8		39		
All	Treatment	99.7	109.9	9.5	153	14.30	.001
	Reg. HS	89.0	100.3		164		
	Treatment	99.7	112.5	17.2	153	33.95	.001
	Alt. HS	93.3	95.3		140		
	Treatment	90.0	117.2	19.2	26	5.92	.025
	Dropout	106.8	98.0		18		

.01). The third-cohort, across-site gain also increased and became statistically significant (gain = 8.0, $F = 9.81$, $p < .005$), as did the combined third-and-fourth cohort, across-site estimate (gain = 4.6, $F = 6.78$, $p < .005$).

The situation is somewhat more positive at posttest time, with two sites showing statistically significant ANCOVA gain estimates for one of the two cohorts as well as for the two-cohort combination. Across sites, the third-cohort and the combined third-and-fourth-cohort gain estimates are statistically significant. This pattern, with the fourth-cohort gain nonsignificant, matches that

Table 28
Treatment Group Raw Score Gains
on the CDI Planning Scale at Posttest Time:
Estimates Derived from Standardized Gain Analyses, Third Cohort

Site	Group	Pretest Mean	Adj. Post-test Mean	Gain	N	F	p
A	Treatment	91.7	128.8	15.6	18	2.88	.05
	Reg. HS	102.1	113.2		39		
	Treatment	91.7	132.9	34.1	18	7.84	.005
	Alt. HS	106.3	98.8		28		
	Treatment	91.7	123.7	22.7	18	2.14	--
	Dropout	97.0	101.0		16		
B	Treatment	102.9	127.6	15.3	45	9.35	.005
	Reg. HS	99.6	112.3		41		
	Treatment	102.9	128.5	10.0	45	2.85	.05
	Alt. HS	100.8	118.5		25		
C	Treatment	91.4	132.1	15.6	18	4.91	.025
	Reg. HS	100.4	116.5		51		
	Treatment	91.4	120.5	11.4	18	.75	--
	Alt. HS	103.1	109.1		8		
All	Treatment	97.8	128.9	14.8	81	15.52	.001
	Reg. HS	104.2	114.1		131		
	Treatment	97.8	127.6	19.1	81	12.80	.001
	Alt. HS	103.6	108.5		61		
	Treatment	91.7	123.7	22.7	18	2.14	--
	Dropout	97.0	101.0		16		

observed in the CDI Planning ANCOVAs. Again, the pattern is attributable to the lessened counselor contact available to fourth-cohort students.

Standardized gain analyses performed on CDI Resources posttest data raised most of the gain estimates but only one nonsignificant ANCOVA estimate attained statistical significance. That was the fourth-cohort estimate at Site A which increased from 6.4 to 10.2 raw score points ($F = 3.29$, $p < .05$).

Table 29
Treatment Group Raw Score Gains
on the CDI Resources Scale at Midtest Time:
Estimates Derived from Covariance Analyses

Site	Cohort	Group	Pretest Mean	Adj. Midtest Mean	Gain	N	F	p
A	III	Treat.	76.9	81.1	-2.3	26	.13	--
		Control	87.5	83.5		13		
	IV	Treat.	79.2	80.9	2.4	29	.14	--
		Control	83.6	78.5		26		
	Comb.	Treat.	78.1	81.2	1.3	55	.08	--
		Control	84.9	79.9		39		
B	III	Treat.	78.8	90.1	- .4	80	.01	--
		Control	83.2	90.5		23		
	IV	Treat.	80.3	90.3	.1	40	.00	--
		Control	79.3	90.3		30		
	Comb.	Treat.	79.3	90.2	- .1	120	.00	--
		Control	81.0	90.3		53		
C	III	Treat.	75.1	85.6	4.8	47	1.19	--
		Control	84.7	80.7		29		
	IV	Treat.	74.4	77.6	.7	52	.01	--
		Control	72.7	76.9		10		
	Comb.	Treat.	74.7	81.5	2.0	99	.28	--
		Control	81.6	79.5		39		
D	III	Treat.	78.5	89.0	7.9	47	4.30	.025
		Control	82.5	81.0		15		
	IV	Treat.	83.8	86.6	- .4	73	.02	--
		Control	85.9	86.9		50		
	Comb.	Treat.	81.7	87.6	2.3	120	1.17	--
		Control	85.1	85.4		65		
All	III	Treat.	77.6	87.5	3.2	200	2.15	--
		Control	84.3	84.3		80		
	IV	Treat.	79.9	84.4	- .1	194	.00	--
		Control	82.5	84.5		116		
	Comb.	Treat.	78.7	86.0	1.7	394	1.21	--
		Control	83.2	84.4		196		

Table 30
Treatment Group Raw Score Gains
on the CDI Resources Scale at Posttest Time:
Estimates Derived from Covariance Analyses

Site	Cohort	Group	Pretest Mean	Adj. Post-test Mean	Gain	N	F	p
A	III	Treat.	74.1	84.4	2.5	18	.22	--
		Control	78.3	81.9		16		
	IV	Treat.	76.8	87.5	6.4	20	1.48	--
		Control	85.4	81.1		17		
	Comb.	Treat.	75.5	85.6	3.6	38	.89	--
		Control	82.0	82.0		33		
B	III	Treat.	78.1	95.9	3.7	45	.60	--
		Control	75.0	92.1		17		
	IV	Treat.	77.9	90.4	6.9	32	3.38	.05
		Control	76.9	83.5		25		
	Comb.	Treat.	78.0	93.6	6.7	77	4.78	.025
		Control	77.8	86.9		42		
C	III	Treat.	76.9	83.8	- .6	18	.01	--
		Control	82.8	84.4		13		
	IV	Treat.	79.2	80.3	-12.4	34	4.53	--
		Control	63.9	92.7		11		
	Comb.	Treat.	78.4	81.0	-8.3	52	3.43	--
		Control	74.1	89.3		24		
D	III	Treat.	80.9	94.9	13.6	30	10.46	.01
		Control	81.9	81.3		14		
	IV	Treat.	82.4	92.8	4.5	62	1.39	--
		Control	84.4	88.3		45		
	Comb.	Treat.	81.8	93.5	6.9	92	5.43	.025
		Control	83.8	86.6		59		
All	III	Treat.	78.0	91.5	5.7	111	3.99	.025
		Control	80.8	85.8		60		
	IV	Treat.	79.9	88.8	2.7	148	1.52	--
		Control	80.4	86.1		98		
	Comb.	Treat.	79.1	90.0	4.1	259	5.40	.025
		Control	80.4	85.9		158		

The standardized gain, comparison group analyses are summarized in Tables 31 and 32. At midtest time the gain estimates are large and statistically significant for both the regular and alternative high school comparisons at Sites B and C. Across-site gain estimates derived from analyses involving these two comparison groups are also significant. The pattern is much the same at posttest time except for Site C where the gain estimates decrease in size and failed to attain statistical significance. Overall, the results of these analyses tend to support those of the ANCOVAs.

Table 31
Treatment Group Raw Score Gains
on the CDI Resources Scale at Midtest Time:
Estimates Derived from Standardized Gain Analyses, Third Cohort

Site	Group	Pretest Mean	Adj. Mid-test Mean	Gain	N	F	p
A	Treatment	76.9	81.1	2.3	26	.20	--
	Reg. HS	77.3	78.7		56		
	Treatment	76.9	82.8	3.0	26	.43	--
	Alt. HS	80.5	79.8		49		
	Treatment	76.9	80.6	5.5	26	.47	--
	Dropout	76.6	75.1		18		
B	Treatment	78.8	91.2	8.7	80	8.80	.005
	Reg. HS	82.3	82.5		53		
	Treatment	78.8	90.9	10.0	80	9.52	.005
	Alt. HS	81.5	80.9		52		
C	Treatment	75.1	90.0	6.7	47	3.16	.05
	Reg. HS	84.4	83.3		55		
	Treatment	75.1	90.1	14.0	47	9.64	.005
	Alt. HS	85.5	76.1		39		
All	Treatment	77.4	88.6	6.5	153	9.28	.005
	Reg. HS	81.3	82.0		164		
	Treatment	77.4	88.8	9.1	153	19.77	.001
	Alt. HS	82.2	79.7		140		
	Treatment	76.9	80.6	5.5	26	.47	--
	Dropout	76.6	75.1		18		

Table 32
Treatment Group Raw Score Gains
on the CDI Resources Scale at Posttest Time:
Estimates Derived from Standardized Gain Analyses, Third Cohort

Site	Group	Pretest Mean	Adj. Post-test Mean	Gain	N	F	p
A	Treatment	74.1	85.3	2.3	18	.16	--
	Reg. HS	75.8	83.0		39		
	Treatment	74.1	83.0	6.0	18	.75	--
	Alt. HS	72.4	77.0		28		
	Treatment	74.1	81.0	11.3	18	1.32	--
	Dropout	68.5	69.7		16		
B	Treatment	78.1	97.5	12.9	45	12.76	.001
	Reg. HS	81.2	84.6		41		
	Treatment	78.1	97.3	14.1	45	9.66	.005
	Alt. HS	82.4	83.2		25		
C	Treatment	76.9	90.0	3.2	18	.45	--
	Reg. HS	86.4	86.8		51		
	Treatment	76.9	84.6	6.3	18	.39	--
	Alt. HS	84.5	78.3		8		
All	Treatment	77.0	93.2	8.2	81	9.27	.005
	Reg. HS	81.6	85.0		131		
	Treatment	77.0	90.4	9.5	81	6.54	.01
	Alt. HS	78.1	80.9		61		
	Treatment	74.1	81.0	11.3	18	1.32	--
	Dropout	68.5	69.7		16		

ANCOVAs performed on scores from the CDI Information scale produced no statistically significant gain estimates at midtest time (see Table 33). The posttest analyses (Table 34) are substantially more positive with 4 of 12 individual-site and all 3 across-site gain estimates attaining statistical significance. Standardized gain analyses increased most of the gain estimates, found statistical significance in one case where the corresponding ANCOVA did not (Site A, fourth cohort; gain = 3.0, $F = 3.33$, $p < .05$), and increased the significance level of two other estimates (Site A, combined, and Site B, fourth cohort).

Table 33
Treatment Group Raw Score Gains
on the CDI Information Scale at Midtest Time:
Estimates Derived from Covariance Analyses

Site	Cohort	Group	Pretest Mean	Adj. Mid-test Mean	Gain	N	F	p
A	III	Treat.	12.2	13.1	1.9	26	1.69	--
		Control	10.8	11.2		13		
	IV	Treat.	9.9	13.3	--	29	--	--
		Control	11.0	13.3		26		
	Comb.	Treat.	11.0	13.3	.7	55	.71	--
		Control	10.9	12.5		39		
B	III	Treat.	12.8	14.3	.1	82	.01	--
		Control	12.5	14.2		25		
	IV	Treat.	12.4	14.1	.1	40	.03	--
		Control	13.2	14.0		30		
	Comb.	Treat.	12.7	14.2	.2	122	.06	--
		Control	12.9	14.1		55		
C	III	Treat.	13.4	13.7	--	47	--	--
		Control	14.2	13.7		29		
	IV	Treat.	12.8	12.9	.1	52	.00	--
		Control	11.5	12.8		10		
	Comb.	Treat.	13.1	13.3	--	99	--	--
		Control	13.5	13.3		39		
D	III	Treat.	13.7	15.0	.5	47	.19	--
		Control	13.6	14.6		15		
	IV	Treat.	12.4	13.2	-1.0	73	4.21	--
		Control	13.4	14.6		50		
	Comb.	Treat.	12.9	13.9	- .7	120	1.71	--
		Control	13.5	14.7		65		
All	III	Treat.	13.1	14.2	.6	202	1 37	--
		Control	13.0	13.6		82		
	IV	Treat.	12.1	13.3	- .7	194	2.38	--
		Control	12.6	14.0		116		
	Comb.	Treat.	12.6	13.8	- .1	396	.14	--
		Control	12.8	13.9		198		

Table 34
Treatment Group Raw Score Gains
on the CDI Information Scale at Posttest Time:
Estimates Derived from Covariance Analyses

Site	Cohort	Group	Pretest Mean	Adj. Post-test Mean	Gain	N	F	p
A	III	Treat.	12.4	14.0	2.1	18	3.04	.05
		Control	11.6	11.9		16		
	IV	Treat.	8.1	13.4	1.5	20	.98	--
		Control	11.2	11.9		17		
	Comb.	Treat.	10.2	13.8	1.9	38	3.90	.05
		Control	11.4	11.9		33		
B	III	Treat.	13.5	16.2	1.5	45	2.73	--
		Control	11.5	14.7		19		
	IV	Treat.	12.3	15.0	1.6	32	3.45	.05
		Control	13.8	13.4		25		
	Comb.	Treat.	13.0	15.7	1.8	77	8.23	.025
		Control	12.8	13.9		44		
C	III	Treat.	12.9	15.3	1.4	18	.94	--
		Control	14.5	13.9		13		
	IV	Treat.	13.8	12.5	.3	34	.03	--
		Control	9.0	12.2		11		
	Comb.	Treat.	13.5	13.4	.3	52	.12	--
		Control	12.0	13.1		24		
D	III	Treat.	14.8	15.4	.6	30	.14	--
		Control	12.9	14.8		14		
	IV	Treat.	12.6	13.4	.8	62	1.02	--
		Control	13.4	12.6		45		
	Comb.	Treat.	13.3	14.0	.9	92	1.60	--
		Control	13.3	13.1		59		
All	III	Treat.	13.6	15.4	1.4	111	4.84	.025
		Control	12.5	14.0		62		
	IV	Treat.	12.2	13.6	1.1	148	4.64	.025
		Control	12.6	12.5		98		
	Comb.	Treat.	12.8	14.4	1.3	259	10.27	.005
		Control	12.6	13.1		160		

It appears that Sites A and B outperformed Sites C and D, but no convincing explanation for this finding occurs to the authors. The fourth-cohort gain estimate is smaller than that for the third cohort, thus continuing the pattern observed with the other two CDI scales. The difference here, however, is small and statistically non-significant.

The standardized gain analyses presented in Tables 35 and 36 closely parallel the corresponding ANCOVAs. None of the resulting gain estimates is statistically significant at midtest time, but approximately half are significant at posttest time. Across sites, the gains at posttest time are also close in size to the estimates derived from the covariance analyses.

Table 35
Treatment Group Raw-Score Gains
on the CDI Information Scale at Midtest Time:
Estimates Derived from Standardized Gain Analyses, Third Cohort

Site	Group	Pretest Mean	Adj. Mid-test Mean	Gain	N	F	p
A	Treatment	12.2	14.6	.5	26	.25	--
	Reg. HS	13.9	14.0		56		
	Treatment	12.2	13.8	.9	26	.88	--
	Alt. HS	12.9	13.0		49		
	Treatment	12.2	14.1	.6	26	.17	--
	Dropout	13.7	13.5		18		
B	Treatment	12.8	14.8	.9	82	1.85	--
	Reg. HS	14.0	13.9		58		
	Treatment	12.8	15.0	.2	82	.07	--
	Alt. HS	14.4	14.8		52		
C	Treatment	13.4	14.8	- .3	47	.11	--
	Reg. HS	15.8	15.1		55		
	Treatment	13.4	14.6	-1.4	47	3.50	--
	Alt. HS	15.8	16.0		39		
All	Treatment	12.9	14.8	.5	155	1.05	--
	Reg. HS	14.6	14.3		164		
	Treatment	12.9	14.6	.0	155	.00	--
	Alt. HS	14.3	14.6		140		
	Treatment	12.2	14.1	.6	26	.17	--
	Dropout	13.7	13.5		18		

Table 36
Treatment Group Raw Score Gains
on the CDI Information Scale at Posttest Time:
Estimates Derived from Standardized Gain Analyses, Third Cohort

Site	Group	Pretest Mean	Adj. Post-test Mean	Gain	N	F	p
A	Treatment	12.4	15.8	- .6	18	.39	--
	Reg. HS	14.9	16.4		39		
	Treatment	12.4	14.9	2.5	18	4.04	.05
	Alt HS	13.6	12.4		28		
	Treatment	12.4	14.7	1.4	18	1.36	--
	Dropout	13.2	13.3		16		
B	Treatment	13.5	16.7	.8	45	1.05	--
	Reg. HS	14.1	15.9		42		
	Treatment	13.5	16.1	.2	45	.05	--
	Alt. HS	12.3	15.9		25		
C	Treatment	12.9	16.4	2.3	18	4.71	.025
	Reg. HS	15.7	14.1		51		
	Treatment	12.9	16.0	.5	18	.14	--
	Alt. HS	16.1	15.5		8		
All	Treatment	13.2	16.6	1.3	81	5.62	.01
	Reg. HS	15.0	15.3		132		
	Treatment	13.2	15.7	1.2	81	3.11	.05
	Alt. HS	13.4	14.		61		
	Treatment	12.4	14.7	1.4	18	1.36	--
	Dropout	13.2	13.3		16		

Self-Esteem Inventory

Tables 37 and 38 present the raw score gains made by second-cohort CIP students on the Coopersmith Self-Esteem Inventory between pre- and midtesting and between pre- and posttesting respectively. At midtest time, two of the individual-site as well as the across-site gain estimates on the Self-Esteem scale are statistically significant. At posttest time, however, there are no significant self-esteem gains.

Table 37
Treatment Group Pre-to-Midtest Raw Score Gains:
Self-Esteem Inventory, Second Cohort

	Pretest Mean	Midtest Mean	Gain	N	t	p
Site A						
Self-Esteem	35.1	39.0	3.9	21	1.64	--
Openness	1.7	2.7	1.0	21	1.92	.05
Site B						
Self-Esteem	33.7	37.9	4.1	38	3.42	.005
Openness	2.6	2.9	.2	38	.71	--
Site C						
Self-Esteem	33.1	34.9	1.8	28	1.24	--
Openness	2.2	2.1	- .1	28	.11	--
Site D						
Self-Esteem	38.3	41.3	3.0	15	2.60	.025
Openness	2.9	3.7	.8	15	1.29	--
All Sites						
Self-Esteem	34.5	37.8	3.3	102	4.17	.001
Openness	2.4	2.8	.4	102	1.96	.025

Table 38
Treatment Group Pre-to-Posttest Raw Score Gains:
Self-Esteem Inventory, Second Cohort

	Pretest Mean	Posttest Mean	Gain	N	t	p
Site A						
Self-Esteem	34.2	38.7	4.6	18	1.54	--
Openness	1.9	2.5	.6	18	1.54	--
Site B						
Self-Esteem	31.1	34.8	3.7	14	1.04	--
Openness	2.1	3.2	1.1	14	2.11	.05
Site C						
Self-Esteem	33.1	31.8	-1.3	9	.48	--
Openness	2.8	3.7	.9	9	1.45	--
Site D						
Self-Esteem	42.0	41.5	- .5	4	.20	--
Openness	3.7	4.2	.5	4	.48	--
All Sites						
Self-Esteem	33.7	36.4	2.7	45	1.55	--
Openness	2.3	3.1	.8	45	3.04	.005

The low midtest gain at Site C is clearly consistent with events at that site. Both implementation and climate were at their lowest point at the time the second cohort was midtested. The larger and statistically significant gains at Sites B and D also make sense in terms of what was happening there. The Site A gain, because of its numerical value, seems inconsistent with the status of implementation there. It must be noted, however, that the gain estimate is not significantly different from zero, a fact that restores consistency between the gain and the site events.

At posttest time it is somewhat surprising that the gain at Site D was not positive and significant. With a sample size of only four, however, such an expectation is unreasonable and the small negative gain shown by those four individuals cannot be taken as any indication of program impact on self-esteem. The small sample size at Site B may also be responsible for the lack of a statistically significant gain.

One individual-site and the across-site Openness gains were statistically significant, both at midtest and at posttest time. This finding, however, appears unrelated to any of the CIP objectives. It may represent no more than the result of repeated exposure to the instrument.

Tables 39 and 40, which include self-esteem gain estimates and related statistics for third- and fourth-cohort CIP participants derived from covariance analyses, present an almost totally negative picture. Although the across-site estimate for third-cohort students at posttest time is significant at the .05 level, only 1 of the other 29 gain estimates was found to be reliably greater than zero.

While these results are not very different from the raw score gains made by second-cohort CIP participants, it seemed that they might be somewhat deflated by a kind of John Henry effect. Since all control group students had been denied access to the program but were mid- and posttested at the CIP facility, it seemed not unlikely that they might distort self-reports in a positive way to cover up the deprivation they felt. With this possibility in mind, a decision was made to examine the raw score gains made by members of the treatment and control groups.

Across sites, the third-cohort treatment group gained 3.5 points, a gain that would almost certainly have been significant with 111 degrees of freedom. The control group, on the other hand, gained 2.7 points. It is not clear whether that control group gain can be attributed to a John Henry effect or whether it stemmed from other causes. Some support for the former hypothesis, however, is afforded by the fact that the regular and alternative high school comparison groups, which comprised students who had not been denied access to the program and who were not tested at the CIP facility, made smaller self-esteem gains than the third-cohort control group (1.4 raw score points in both instances).

In any case, the control group gain enters into the covariance calculations and reduces both the size and the significance level of the ANCOVA gain estimate. At Site A, the situation is even worse. Although the treatment group gained 4.1 points, the control group gained 5.4. A similar, although less dramatic, pattern is seen in the fourth-cohort data. There the treatment group gained 2.3 points while the control group gained 1.4. At Site B the treatment group made a gain of 4.1 points but it was largely offset by the 3.4 points gained by the control group.

One interesting finding that shows up in these analyses is that the fourth cohort made smaller gains than the third cohort. If one assumes that improved self-esteem is at least partially a counseling outcome, then this finding is consistent with the reduced amount of counseling available to fourth-cohort students--a situation that apparently influenced other scores as well.

Tables 41 and 42 summarize the results of the standardized gain analyses involving the three comparison groups. None of the gain estimates is significant at midtest time (Table 41) but two individual-site and two across-site estimates are significant at the .05 level at posttest time (Table 42). It is also noteworthy that the two significant individual-site gain estimates occur at Site B,

Table 39
Treatment Group Raw Score Gains
on the Self-Esteem Scale at Midtest Time:
Estimates Derived from Covariance Analyses

Site	Cohort	Group	Pretest Mean	Adj. Mid-test Mean	Gain	N	F	p
A	III	Treat.	34.0	33.5	- .8	28	.08	--
		Control	32.4	34.3		11		
	IV	Treat.	35.8	39.2	1.1	29	.50	--
		Control	35.2	38.0		26		
	Comb.	Treat.	34.9	36.5	- .4	57	.06	--
		Control	34.4	36.8		37		
B	III	Treat.	36.5	38.3	1.5	81	1.34	--
		Control	36.0	36.8		23		
	IV	Treat.	34.7	37.4	.1	40	.01	--
		Control	35.3	37.2		30		
	Comb.	Treat.	35.9	38.0	.8	121	.91	--
		Control	35.6	37.1		53		
C	III	Treat.	35.0	36.6	.6	47	.16	--
		Control	36.2	36.0		28		
	IV	Treat.	34.2	36.6	1.3	52	.41	--
		Control	35.9	35.3		10		
	Comb.	Treat.	34.6	36.6	.9	99	.65	--
		Control	36.1	35.7		38		
D	III	Treat.	35.3	39.1	1.9	48	1.82	--
		Control	37.5	37.2		14		
	IV	Treat.	36.0	38.1	.4	75	.14	--
		Control	37.7	37.8		49		
	Com'	Treat.	35.7	38.5	1.0	123	1.33	--
		Control	37.7	37.6		63		
All	III	Treat.	35.5	37.4	1.1	204	1.87	--
		Control	35.8	36.3		76		
	IV	Treat.	35.2	37.7	.2	196	.13	--
		Control	36.4	37.5		115		
	Comb.	Treat.	35.4	37.6	.6	400	1.18	--
		Control	36.2	37.0		191		

Table 40
Treatment Group Raw Score Gains
on the Self-Esteem Scale at Posttest Time:
Estimates Derived from Covariance Analyses

Site	Cohort	Group	Pretest Mean	Adj. Post-test Mean	Gain	N	F	p
A	III	Treat.	34.5	37.8	1.4	19	.39	--
		Control	30.1	36.4		16		
	IV	Treat.	36.1	40.2	2.7	20	1.38	--
		Control	35.7	37.5		17		
	Comb.	Treat.	35.3	39.1	2.1	39	1.83	--
		Control	33.0	37.0		33		
B	III	Treat.	36.9	40.2	2.9	45	2.63	--
		Control	33.9	37.3		15		
	IV	Treat.	36.2	39.5	1.9	31	2.39	--
		Control	33.2	37.6		25		
	Comb.	Treat.	36.6	39.8	2.3	76	4.87	.025
		Control	33.5	37.5		40		
C	III	Treat.	34.8	36.6	.3	18	.01	--
		Control	36.5	36.3		13		
	IV	Treat.	34.6	36.8	1.5	34	.64	--
		Control	35.7	35.3		11		
	Comb.	Treat.	34.7	36.8	1.0	52	.46	--
		Control	36.1	35.8		24		
D	III	Treat.	35.3	39.9	1.2	30	.30	--
		Control	37.8	38.7		14		
	IV	Treat.	35.8	37.5	.3	62	.05	--
		Control	38.2	37.8		44		
	Comb.	Treat.	35.7	38.2	.1	92	.02	--
		Control	38.1	38.1		58		
All	III	Treat.	35.7	39.0	1.7	112	2.81	.05
		Control	34.3	37.3		58		
	IV	Treat.	35.7	38.1	.7	147	.59	--
		Control	36.2	37.4		97		
	Comb.	Treat.	35.7	38.5	1.1	259	2.53	--
		Control	35.5	37.4		155		

Table 41
Treatment Group Raw Score Gains
on the Self-Esteem Scale at Midtest Time:
Estimates Derived from Standardized Gain Analyses, Third Cohort

Site	Group	Pretest Mean	Adj. Mid-test Mean	Gain	N	F	p
A	Treatment	34.0	34.6	- .1	28	.00	--
	Reg. HS	35.2	34.7		55		
	Treatment	34.0	33.2	1.3	28	.54	--
	Alt. HS	32.9	34.5		43		
	Treatment	34.0	34.6	- .6	28	.08	--
	Dropout	36.4	35.2		19		
B	Treatment	36.5	37.4	1.6	81	2.19	--
	Reg. HS	34.0	38.8		52		
	Treatment	36.5	39.0	1.8	81	2.62	--
	Alt. HS	38.1	37.1		55		
C	Treatment	35.0	37.5	1.0	47	.61	--
	Reg. HS	36.8	36.5		54		
	Treatment	35.0	36.8	1.2	47	.89	--
	Alt. HS	35.8	35.6		39		
All	Treatment	35.6	36.8	.9	156	1.57	--
	Reg. HS	35.4	35.8		161		
	Treatment	35.6	37.0	.9	156	1.25	--
	Alt. HS	35.8	36.1		137		
	Treatment	34.0	34.6	- .6	28	.08	--
	Dropout	36.4	35.2		19		

the only site where implementation was nearly ideal throughout the entire year between pre- and posttesting of third-cohort students.

The standardized gain analyses produced substantially more positive results than the ANCOVAs. As suggested earlier, this difference tends to support the hypothesis that control group students may have biased their reports of self-esteem in a positive direction because they had been denied entry into the program. It seems likely, in view of this possibility, that the standardized gain analyses provide more valid estimates of program impact on self-esteem than the covariance analyses.

Table 42
Treatment Group Raw Score Gains
on the Self-Esteem Scale at Posttest Time:
Estimates Derived from Standardized Gain Analyses, Third Cohort

Site	Group	Pretest Mean	Adj. Post-test Mean	Gain	N	F	p
A	Treatment	34.5	38.0	1.0	19	.23	--
	Reg. HS	33.4	37.0		39		
	Treatment	34.5	37.2	2.6	19	1.41	--
	Alt. HS	31.6	34.5		28		
	Treatment	34.5	38.6	1.4	19	.42	--
	Dropout	34.4	37.2		16		
B	Treatment	36.9	39.1	2.5	45	3.18	.05
	Reg. HS	34.0	36.6		41		
	Treatment	36.9	40.0	3.4	45	3.41	.05
	Alt. HS	35.7	36.6		26		
C	Treatment	34.8	38.2	1.8	18	.75	--
	Reg. HS	36.9	36.3		50		
	Treatment	34.8	36.5	-1.3	18	.18	--
	Alt. HS	35.2	37.8		8		
All	Treatment	35.9	38.5	1.8	82	3.22	.05
	Reg. HS	34.9	36.7		130		
	Treatment	35.9	38.5	2.0	82	3.19	.05
	Alt. HS	34.3	36.5		62		
	Treatment	34.5	38.6	1.4	19	.42	--
	Dropout	34.4	37.2		16		

Tables 43 through 46 summarize the results of the covariance and standardized gain analyses performed on Coopersmith Openness scores. None of the across-site analyses shows a significant gain estimate and only 5 of the 50 individual-site estimates are statistically significant. (Two-tailed tests were used in these analyses as there was no reason to predict that the program treatment would either raise or lower scores on this scale.)

The nonsignificance and apparent irrelevance of these gain estimates to program goals suggests that no further attempts at interpretation be made.

Table 43
Treatment Group Raw Score Gains
on the Openness Scale at Midtest Time:
Estimates Derived from Covariance Analyses

Site	Cohort	Group	Pretest Mean	Adj. Mid-test Mean	Gain	N	F	p
A	III	Treat.	2.6	3.8	.9	28	2.35	--
		Control	3.2	2.9		12		
	IV	Treat.	2.9	3.2	.7	29	2.52	--
		Control	2.2	2.5		26		
	Comb.	Treat.	2.7	3.5	.8	57	5.29	.025*
		Control	2.5	2.7		38		
B	III	Treat.	2.8	2.4	- .8	81	4.84	.05 *
		Control	3.4	3.3		22		
	IV	Treat.	2.9	2.6	.1	40	.01	--
		Control	2.6	2.6		30		
	Comb.	Treat.	2.8	2.5	- .4	121	2.07	--
		Control	2.9	2.9		52		
C	III	Treat.	2.1	2.9	- .2	47	.41	--
		Control	2.8	3.1		28		
	IV	Treat.	2.7	3.2	.4	52	.48	--
		Control	3.4	2.8		10		
	Comb.	Treat.	2.4	3.0	- .1	99	.04	--
		Control	3.0	3.1		38		
D	III	Treat.	2.6	2.8	.5	48	1.02	--
		Control	2.1	2.3		14		
	IV	Treat.	2.9	2.4	- .5	75	2.83	--
		Control	2.8	2.9		49		
	Comb.	Treat.	2.8	2.6	- .2	123	.59	--
		Control	2.6	2.8		63		
All	III	Treat.	2.6	2.8	- .2	204	1.08	--
		Control	2.9	3.0		76		
	IV	Treat.	2.8	2.8	.0	196	.01	--
		Control	2.6	2.7		115		
	Comb.	Treat.	2.7	2.8	- .1	400	.19	--
		Control	2.7	2.8		191		

*Two-tailed probability

Table 44
Treatment Group Raw Score Gains
on the Openness Scale at Posttest Time:
Estimates Derived from Covariance Analyses

Site	Cohort	Group	Pretest Mean	Adj. Post-test Mean	Gain	N	F	p
A	III	Treat.	3.2	3.0	1.3	19	4.63	.05
		Control	2.2	1.7		15		
	IV	Treat.	3.2	2.9	- .3	20	.16	--
		Control	1.9	3.2		17		
	Comb.	Treat.	3.2	3.0	.6	39	1.29	--
		Control	2.0	2.4		32		
B	III	Treat.	2.6	2.8	.1	45	.05	--
		Control	2.7	2.7		15		
	IV	Treat.	2.9	3.3	.9	31	4.51	.025
		Control	2.6	2.4		25		
	Comb.	Treat.	2.7	3.0	.5	76	2.72	--
		Control	2.6	2.5		40		
C	III	Treat.	1.8	2.5	- .3	18	.20	--
		Control	2.5	2.8		13		
	IV	Treat.	2.9	3.2	- .6	34	1.06	--
		Control	3.6	3.8		11		
	Comb.	Treat.	2.5	3.0	- .2	52	.49	--
		Control	3.0	3.2		24		
D	III	Treat.	2.9	3.2	.6	30	.91	--
		Control	2.1	2.6		12		
	IV	Treat.	2.9	2.9	.2	62	.38	--
		Control	2.9	3.1		44		
	Comb.	Treat.	2.9	3.0	--	92	--	--
		Control	2.8	3.0		56		
All	III	Treat.	2.6	2.9	.5	112	2.57	--
		Control	2.4	2.4		55		
	IV	Treat.	2.9	3.1	.1	147	.04	--
		Control	2.7	3.0		97		
	Comb.	Treat.	2.8	3.0	.2	259	1.18	--
		Control	2.6	2.8		152		

Table 45
Treatment Group Raw Score Gains
on the Openness Scale at Midtest Time:
Estimates Derived from Standardized Gain Analyses, Third Cohort

Site	Group	Pretest Mean	Adj. Mid-test Mean	Gain	N	F	p
A	Treatment	2.6	3.6	.9	28	5.65	.05
	Reg. HS	2.3	2.6		55		
	Treatment	2.6	3.8	- .1	28	.01	--
	Alt. HS	2.6	3.8		43		
	Treatment	2.6	3.8	.0	28	.00	--
	Dropout	2.6	3.8		19		
B	Treatment	2.8	2.2	- .6	81	3.49	--
	Reg. HS	2.4	2.8		52		
	Treatment	2.8	2.3	.0	81	.01	--
	Alt. HS	2.6	2.4		55		
C	Treatment	2.1	2.9	.3	47	.73	--
	Reg. HS	2.4	2.6		54		
	Treatment	2.1	2.6	.6	47	3.52	--
	Alt. HS	1.9	2.0		39		
All	Treatment	2.5	2.6	- .1	156	.09	--
	Reg. HS	2.4	2.7		161		
	Treatment	2.5	2.6	- .1	156	.26	--
	Alt. HS	2.4	2.8		137		
	Treatment	2.6	3.8	.0	28	.00	--
	Dropout	2.6	3.8		19		

Table 46
Treatment Group Raw Score Gains
on the Openness Scale at Posttest Time:
Estimates Derived from Standardized Gain Analyses, Third Cohort

Site	Group	Pretest Mean	Adj. Post-test Mean	Gain	N	F	p
A	Treatment	3.2	2.8	.4	19	.47	--
	Reg. HS	2.5	2.4		39		
	Treatment	3.2	2.4	- .3	19	.36	--
	Alt. HS	2.2	2.7		28		
	Treatment	3.2	2.9	.8	19	1.73	--
	Dropout	2.6	2.1		16		
B	Treatment	2.6	2.8	.3	45	.06	--
	Reg. HS	2.5	2.5		41		
	Treatment	2.6	2.6	.1	45	.42	--
	Alt. HS	2.3	2.5		26		
C	Treatment	1.8	2.2	- .4	18	.25	--
	Reg. HS	1.5	2.6		50		
	Treatment	1.8	2.8	.3	18	.81	--
	Alt. HS	2.3	2.5		8		
All	Treatment	2.6	2.6	--	82	--	--
	Reg. HS	2.3	2.6		130		
	Treatment	2.6	2.7	.3	82	.81	--
	Alt. HS	2.4	2.4		62		
	Treatment	3.2	2.9	.8	19	1.73	--
	Dropout	2.6	2.1		16		

Internal-External Scale

The results of pre-to-midtest and pre-to-posttest raw score gain analyses for second-cohort CIP participants are summarized, respectively, in Tables 47 and 48. None of the pre-to-midtest gains and only one of the pre-to-posttest gain is statistically significant. The sample sizes for the individual-site, pre-to-posttest analyses are all quite small and account, in large measure, for the negative results. The larger sample size for the across-site gain was responsible for the significant *t*.

Table 47
Treatment Group Pre-to-Midtest Raw Score Gains:
Internal-External Scale, Second Cohort

	Pretest Mean	Midtest Mean	Gain	N	*t*	p
Site A	15.8	17.0	1.3	22	1.30	--
Site B	15.8	15.8	.0	40	.04	--
Site C	15.4	14.1	-1.3	26	1.72	--
Site D	15.0	14.9	- .1	15	.19	--
All Sites	15.6	15.5	- .1	103	.19	--

Table 48
Treatment Group Pre-to-Posttest Raw Score Gains:
Internal-External Scale, Second Cohort

	Pretest Mean	Posttest Mean	Gain	N	*t*	p
Site A	15.9	17.9	1.9	18	1.69	--
Site B	15.0	16.5	1.5	15	1.02	--
Site C	13.2	14.1	.9	9	.57	--
Site D	17.0	18.6	1.6	5	.83	--
All Sites	15.2	16.8	1.6	47	2.18	.025

Tables 49 and 50 summarize the results of the ANCOVAs. Only one of the individual-site gain estimates is statistically significant and none of the across-site analyses shows a significant F. It is hypothesized that the same forces might be operating here as appeared to operate in the case of the Self-Esteem scale--in other words, that members of the control group might deliberately distort their responses in order to appear in a more favorable light. The data, however, did not offer strong support for this hypothesis.

In terms of raw scores, the third-cohort treatment group shows a pre-to-posttest gain of 1.1 which is statistically significant ($t = 2.87$, $df = 112$, $p < .01$). The control group has a gain of .6 raw score points, which is nonsignificant but large enough to prevent the ANCOVA from showing a significant gain. Had data from the fourth cohort presented a similar picture, a plausible case could have been made for biased self-reporting. In fact, however, the fourth-cohort control group's mean posttest score is lower than its pretest score (although not significantly). While the gain made by the treatment group is only .2 raw score points, the control group's performance served to inflate the ANCOVA estimate yielding a value of .4 points. This finding appeared to negate the John Henry hypothesis.

Tables 51 and 52 summarize the results of the standardized gain analyses. Although one individual-site and one across-site gain estimate are significant at posttest time, the picture suggests that the CIP does not strongly or consistently affect locus of control. If there is any effect, it is slow to develop. None of the gains from any of the analyses is significant at midtest time. Neither are any of the fourth-cohort gains significant after nine months (the pre-to-posttest interval for that cohort).

Table 49
Treatment Group Raw Score Gains
on the Internal-External Scale at Midtest Time:
Estimates Derived from Covariance Analyses

Site	Cohort	Group	Pretest Mean	Adj. Mid-test Mean	Gain	N	F	p
A	III	Treat.	16.6	15.4	-1.4	28	9.25	--
		Control	14.8	17.7		12		
	IV	Treat.	15.2	16.4	.9	28	1.25	--
		Control	15.5	15.5		26		
	Comb.	Treat.	15.9	15.9	- .3	56	.33	--
		Control	15.2	16.2		38		
B	III	Treat.	15.8	16.1	.2	80	.07	--
		Control	16.2	15.9		24		
	IV	Treat.	15.1	15.6	.1	40	.02	--
		Control	15.5	15.4		28		
	Comb.	Treat.	15.6	15.9	.2	120	.17	--
		Control	15.8	15.7		52		
C	III	Treat.	15.5	16.4	- .4	46	.38	--
		Control	15.6	16.8		28		
	IV	Treat.	14.8	14.5	- .6	52	.27	--
		Control	17.5	15.1		10		
	Comb.	Treat.	15.2	15.4	- .9	98	2.30	--
		Control	16.1	16.3		38		
D	III	Treat.	15.8	16.8	.3	46	.11	--
		Control	16.1	16.5		15		
	IV	Treat.	15.2	15.7	- .6	65	.69	--
		Control	16.1	16.4		45		
	Comb.	Treat.	15.4	16.2	- .2	111	.21	--
		Control	16.1	16.4		60		
All	III	Treat.	15.9	16.2	- .5	200	1.48	--
		Control	15.8	16.7		79		
	IV	Treat.	15.1	15.5	- .3	185	.31	--
		Control	15.9	15.8		109		
	Comb.	Treat.	15.5	15.8	- .3	385	1.42	--
		Control	15.8	16.2		188		

Table 50
Treatment Group Raw Score Gains
on the Internal-External Scale at Posttest Time:
Estimates Derived from Covariance Analyses

Site	Cohort	Group	Pretest Mean	Adj. Post-test Mean	Gain	N	F	p
A	III	Treat.	17.0	16.8	.9	19	.79	--
		Control	14.4	15.9		16		
	IV	Treat.	15.0	16.3	.1	20	.01	--
		Control	14.9	16.2		17		
	Comb.	Treat.	15.9	16.6	.6	39	.54	--
		Control	14.6	16.0		33		
B	III	Treat.	15.5	16.8	1.7	46	3.20	.05
		Control	15.5	15.1		16		
	IV	Treat.	15.1	15.3	- .6	31	.45	--
		Control	15.3	15.9		25		
	Comb.	Treat.	15.4	16.2	.6	77	1.07	--
		Control	15.4	15.6		41		
C	III	Treat.	15.9	17.2	- .5	18	.23	--
		Control	15.8	17.7		13		
	IV	Treat.	14.9	15.7	1.1	34	1.03	--
		Control	16.2	14.6		11		
	Comb.	Treat.	15.2	16.2	- .2	52	.05	--
		Control	16.0	16.4		24		
D	III	Treat.	15.6	16.4	.5	30	.16	--
		Control	15.4	15.9		14		
	IV	Treat.	15.1	14.9	.1	54	.02	--
		Control	16.6	14.8		45		
	Comb.	Treat.	15.3	15.3	.2	84	.10	--
		Control	16.3	15.1		59		
All	III	Treat.	15.8	16.8	.8	113	2.37	--
		Control	15.2	16.0		59		
	IV	Treat.	15.0	15.4	--	139	--	--
		Control	15.9	15.4		98		
	Comb.	Treat.	15.4	16.0	.4	252	.83	--
		Control	15.7	15.6		157		

Table 51
Treatment Group Raw Score Gains
on the Internal-External Scale at Midtest Time:
Estimates Derived from Standardized Gain Analyses, Third Cohort

Site	Group	Pretest Mean	Adj. Mid-test Mean	Gain	N	F	p
A	Treatment	16.6	15.4	- .3	28	1.41	--
	Reg. HS	16.5	16.2		55		
	Treatment	16.6	14.3	-1.9	28	5.22	--
	Alt. HS	14.6	16.2		48		
	Treatment	16.6	15.3	- .9	28	.46	--
	Dropout	16.3	16.2		17		
B	Treatment	15.8	16.1	.4	80	.34	--
	Reg. HS	15.9	15.6		52		
	Treatment	15.8	16.5	.1	80	.01	--
	Alt. HS	16.7	16.4		51		
C	Treatment	15.5	16.4	.0	46	.00	--
	Reg. HS	15.4	16.4		53		
	Treatment	15.5	16.7	.4	46	.14	--
	Alt. HS	16.0	16.2		38		
All	Treatment	15.9	16.1	.0	154	.01	--
	Reg. HS	16.0	16.0		160		
	Treatment	15.9	16.0	- .4	154	1.03	--
	Alt. HS	15.8	16.4		137		
	Treatment	16.6	15.3	- .9	28	.46	--
	Dropout	16.3	16.2		17		

Table 52
Treatment Group Raw Score Gains
on the Internal-External Scale at Posttest Time:
Estimates Derived from Standardized Gain Analyses, Third Cohort

Site	Group	Pretest Mean	Adj. Post-test Mean	Gain	N	F	p
A	Treatment	17.0	17.2	1.2	19	1.92	--
	Reg. HS	16.6	16.0		39		
	Treatment	17.0	15.7	-1.4	19	2.05	--
	Alt. HS	13.0	17.1		27		
	Treatment	17.0	17.8	1.7	19	2.45	--
	Dropout	17.8	16.1		14		
B	Treatment	15.5	17.2	1.2	46	2.16	--
	Reg. HS	16.4	16.1		41		
	Treatment	15.5	17.2	2.5	46	5.28	.025
	Alt. HS	16.5	14.7		23		
C	Treatment	15.9	16.9	- 1.8	18	2.47	--
	Reg. HS	15.5	15.0		50		
	Treatment	15.9	17.5	- .4	18	.06	--
	Alt. HS	16.9	17.9		8		
All	Treatment	15.9	17.2	1.6	83	8.63	.005
	Reg. HS	16.1	15.6		130		
	Treatment	15.9	16.7	.1	83	.01	--
	Alt. HS	14.9	16.6		58		
	Treatment	17.0	17.8	1.7	19	2.45	--
	Dropout	17.8	16.1		14		

Follow-Up Outcomes

All of the analyses performed on the follow-up data involve contrasts between the treated and the untreated portions of the treatment group. In those situations where control groups were available, their data were contrasted with those of the total treatment group as well as with those of the treated subgroup. Comparisons between control and treatment groups are less subject to bias resulting from (possibly) differential attrition and are therefore more credible. On the other hand, because members of the untreated subgroup received little or no treatment, the size of the treatment effect is necessarily diminished. Comparisons between the control group and the treated subgroup can be expected to show larger differences, but the possibility that these differences result from self-selection rather than from the treatment is also more plausible.

All of the follow-up data were analyzed using Chi Square techniques. Most of them involved 2 x 2 tables where, for example, the numbers of employed and unemployed youths from treatment and control groups were tallied and compared.

Table 53 presents, by site, cohort, and group, the numbers of students about whom it was possible to obtain some information. For the treatment and control groups (but not for the treated and untreated subgroups), these numbers are also expressed as percentages of the corresponding total groups pretested. As can be seen, it was possible to obtain a much higher percentage of follow-up returns than either mid- or posttest scores. Overall, the first follow-up return percentage was 73% for the combined treatment groups, and 76% for the combined control groups. The corresponding figures for the second follow-up were 76% and 72%.

Site B had the highest return rate for both follow-ups while Site C had the lowest for the first follow-up but was tied with Sites A and D for the second. These individual-site return rates are thought to reflect both the difficulty in locating the students (due to their mobility, for example) and the resourcefulness and zeal of the site assistants. Unfortunately, it is not possible to separate out the relative contribution of these influences.

It should be pointed out that not all of the follow-up data are highly reliable. Where direct contact with the students in question proved impossible, we attempted to gain information from friends, relatives, school records, and other sources. Occasionally, different sources would yield contradictory information about a single individual. One CIP intern, for example, was reported as dropped out and unemployed by a relative when, in fact, he had graduated from the CIP and was enrolled as a full-time student in college. We sorted out such conflicting stories as carefully as we could, but some errors almost certainly remain in the data.

Table 53

Return Rates for the First and Second Follow-Ups by Site, Cohort, and Group

Cohort	Group	Site A		Site B		Site C		Site D	
		1st	2nd	1st	2nd	1st	2nd	1st	2nd
II	Treated	29	30	52	49	25	34	22	20
	Untreated	20	18	12	13	0	3	18	17
	Total	49	48	64	62	25	37	40	37
		(75%)	(74%)	(84%)	(82%)	(51%)	(76%)	(60%)	(55%)
III	Treated	53	50	92	89	30	59	67	77
	Untreated	34	29	13	12	15	27	14	21
	Total	87	79	105	101	45	86	81	98
		(81%)	(73%)	(87%)	(83%)	(38%)	(72%)	(69%)	(82%)
	Control	41	41	49	44	27	38	31	38
		(75%)	(75%)	(82%)	(73%)	(50%)	(70%)	(56%)	(69%)
IV	Treated	52		51		51		95	
	Untreated	47		16		8		22	
	Total	99		67		59		117	
		(98%)		(89%)		(89%)		(66%)	
	Control	46		58		22		95	
		(84%)		(78%)		(76%)		(90%)	
All	Treatment	235	127	236	163	129	123	238	135
		(86%)	(73%)	(87%)	(83%)	(53%)	(73%)	(66%)	(73%)
All	Control	87	41	107	44	49	38	126	38
		(79%)	(75%)	(80%)	(73%)	(59%)	(70%)	(78%)	(69%)

Tables 54 and 55 present statistics relevant to the high school status of second-cohort CIP interns. Across sites, at the time of the first follow-up, two-thirds of the treated group have graduated from high school, were currently enrolled, or had received a GED, while two-thirds of the untreated group have dropped out prior to graduation and have not received a GED. (There were no control group for the second cohort.) At the time of the second follow-up, the results are only slightly less dramatic with two-thirds falling to 63% in the case of the treated group and to 61% in the case of the untreated group. The overall results of both follow-ups are highly significant ($p < .0025$ in both cases).

Table 54

High School Status of Treated and Untreated Group Members: First Follow-Up, Second Cohort

Site	Group	% Grad., GED, or Enrolled	% Dropped Out	Sample Size
A	Treated	69%	31%	29
	Untreated	50%	50%	20
B	Treated	75%	25%	52
	Untreated	50%	50%	12
C	Treated	52%	48%	25
	Untreated	-	-	0
D	Treated	64%	36%	22
	Untreated	6%	94%	18
All	Treated	67%	33%	128
	Untreated	34%	66%	50

Table 55

High School Status of Treated and Untreated Group Members: Second Follow-Up, Second Cohort

Site	Group	% Grad., GED, or Enrolled	% Dropped Out	Sample Size
A	Treated	60%	40%	30
	Untreated	56%	44%	18
B	Treated	71%	29%	49
	Untreated	54%	46%	13
C	Treated	56%	44%	34
	Untreated	33%	67%	3
D	Treated	60%	40%	20
	Untreated	12%	88%	17
All	Treated	63%	37%	133
	Untreated	39%	61%	51

The individual-site findings are most dramatic at Site D where the data suggest that very few of those who did not enroll in the CIP, or who dropped out shortly after enrollment, returned to school or entered GED programs. A partial explanation for this fact is that all of the second-cohort interns at Site D had previously dropped out of school. Apparently, their disenchantment with "the system" continued.

For the treated group, the results were most favorable at Site B, a finding that is consistent with the state of implementation at that site at that time. The individual-site Chi Squares (both first and second follow-ups) were only significant at Site D, however, and primarily because of the high dropout rate in the untreated group.

Tables 56 and 57 present data on the high school status of third-cohort treated, untreated, and control group members. Results from the first follow-up look much like the corresponding second-cohort findings as far as the treated and untreated subgroups are concerned. Across sites approximately 60% of the treated subgroup members have graduated from high school, are currently enrolled, or have earned a GED. Only 40% of the untreated subgroup fall into this category. The control group percentages are approximately half way between those of the treated and untreated subgroups. The treated and untreated subgroups are significantly different (Chi Square = 10.00, $p < .01$) but neither the treatment group nor the treated subgroup is significantly different from the control group. At Site C, however, the treatment group is significantly superior to the control group (Chi Square = 4.18, $p < .05$).

At the time of the second follow-up, the treated and untreated subgroups remain significantly different (Chi Square = 3.9, $p < .05$), but the difference is somewhat smaller than at the time of the first follow-up. The results at Site C continue to favor the treatment over the control group (Chi Square = 4.14. $p < .05$).

At Site A, the control group has a larger percentage of students who have graduated from high school, are currently enrolled, or have obtained a GED than any of the other groups at any of the other sites. This unexpected finding may reflect the fact that Site A had some 20 other alternative programs readily available to students who were having difficulty in high school. In any case, it has an important effect on the overall results. When Site A data are removed, the composite treated subgroup (from the other three sites) has a significantly better high school performance record than the control group (Chi Square = 4.23, $p < .05$).

Table 56

High School Status of Treated, Untreated, and Control Group Members: First Follow-Up, Third Cohort

Site	Group	% Grad., GED, or Enrolled	% Dropped Out	Sample Size
A	Treated	57%	43%	53
	Untreated	38%	62%	34
	Control	61%	39%	41
B	Treated	64%	36%	91
	Untreated	62%	38%	13
	Control	55%	45%	49
C	Treated	57%	43%	30
	Untreated	47%	53%	15
	Control	28%	72%	25
D	Treated	58%	42%	67
	Untreated	14%	86%	14
	Control	53%	47%	30
All	Treated	60%	40%	241
	Untreated	39%	61%	76
	Control	52%	48%	145

Table 57

High School Status of Treated, Untreated, and Control Group Members: Second Follow-Up, Third Cohort

Site	Group	% Grad., GED, or Enrolled	% Dropped Out	Sample Size
A	Treated	46%	54%	50
	Untreated	45%	55%	29
	Control	63%	37%	41
B	Treated	54%	46%	89
	Untreated	42%	58%	12
	Control	52%	48%	44
C	Treated	44%	56%	59
	Untreated	33%	67%	27
	Control	22%	78%	37
D	Treated	49%	51%	77
	Untreated	29%	71%	21
	Control	38%	62%	37
All	Treated	49%	51%	275
	Untreated	37%	63%	89
	Control	45%	55%	159

Table 58 presents high school status information for the fourth cohort at the time of its first (and only) follow-up. Across sites, both the treated versus control and the treatment versus control comparisons are statistically significant (Chi Squares = 18.05 and 13.40 respectively, $p < .001$ in both cases). In both cases, these findings are largely attributable to Site D where 80% of the treated subgroup have graduated from high school, are currently enrolled, or have obtained a GED. This finding, of course, is consistent with the full operational status and positive climate that had emerged at Site D by the time the fourth cohort enrolled.

The control group at Site A continues to present an unexpectedly positive picture with respect to high school status. While it is not surprising that the treatment group shows up as it does (given the state of program implementation at Site A), the control group percentages for Site A are significantly more favorable than those at the other three sites combined (Chi Square = 8.65, $p < .01$).

Table 58

High School Status of Treated, Untreated, and Control Group Members: First Follow-Up, Fourth Cohort

Site	Group	% Grad., GED, or Enrolled	% Dropped Out	Sample Size
A	Treated	62%	38%	52
	Untreated	45%	55%	47
	Control	63%	37%	46
B	Treated	47%	53%	51
	Untreated	62%	38%	16
	Control	52%	48%	58
C	Treated	51%	49%	51
	Untreated	-	100%	8
	Control	32%	68%	22
D	Treated	80%	20%	95
	Untreated	68%	32%	22
	Control	33%	67%	95
All	Treated	63%	37%	249
	Untreated	49%	51%	93
	Control	44%	56%	221

Tables 59 and 60 summarize second-cohort data from the first and second follow-ups. The comparisons are between those who are either enrolled in some type of school program (high school, college, GED, or vocational) or employed (full- or part-time) and those who are neither in school nor employed. At the time of the first follow-up, there are significantly more members of the across-site treated subgroup than of the untreated subgroup who are either in school or employed (Chi Square = 6.66, $p < .01$). Six months later, however, the relationship is no longer significant. In almost every instance, the status of the untreated subgroup is shown to improve while the status of the treated subgroup is shown to deteriorate.

Table 59

School/Employment Status of Treated and Untreated Group Members: First Follow-Up, Second Cohort

Site	Group	% in School or Employed	% Not in School and Unemployed	Sample Size
A	Treated	62%	38%	29
	Untreated	40%	60%	20
B	Treated	60%	40%	52
	Untreated	50%	50%	12
C	Treated	56%	44%	25
	Untreated	-	-	-
D	Treated	82%	18%	22
	Untreated	39%	61%	18
All	Treated	63%	37%	128
	Untreated	42%	58%	50

Table 60

School/Employment Status of Treated and Untreated Group Members: Second Follow-Up, Second Cohort

Site	Group	% in School or Employed	% Not in School and Unemployed	Sample Size
A	Treated	57%	43%	30
	Untreated	44%	56%	18
B	Treated	51%	49%	49
	Untreated	77%	23%	13
C	Treated	68%	32%	34
	Untreated	33%	67%	3
D	Treated	65%	35%	20
	Untreated	47%	53%	17
All	Treated	59%	41%	133
	Untreated	53%	47%	51

The most dramatic difference between treated and untreated subgroups at the time of the first follow-up occurs at Site D. This finding is somewhat surprising in view of the fact that Site D was not functioning well early in the demonstration period. On the other hand, all of the second-cohort interns at Site D were dropouts and most of those who stayed long enough to be counted as treated remained in the program for a long time since they needed many credits to graduate. Most were still there when the program was turned around.

Site D also shows the largest change from the first to the second follow-up. Most of this change, however, can be traced to five individuals who were employed full-time when the first follow-up was completed but who were unemployed six months later. Part of this reduction can be attributed to the fact that more students are employed full-time during the summer (when the first follow-up was undertaken) than during the school year (35% vs. 29% across all sites). Perhaps more important, however, is the fact that the employment situation was quite good at Site D when the first follow-up was undertaken and quite bad six months later.

Tables 61 and 62 present the school/employment status data for the third cohort treated, untreated, and control groups. As was the case with the second cohort, the across-site treated group is significantly better off than the untreated group at the time of the first follow-up (Chi Square = 5.62, $p < .025$). The difference, however, becomes nonsignificant by the time of the second. None of the

treated-versus-control or treatment-versus-control comparisons is statistically significant either at individual sites or across sites on the first or the second follow-up.

Table 61

School/Employment Status of Treated, Untreated, and Control Group Members: First Follow-Up, Third Cohort

Site	Group	% in School or Employed	% Not in School and Unemployed	Sample Size
A	Treated	66%	34%	53
	Untreated	50%	50%	34
	Control	71%	29%	41
B	Treated	72%	28%	92
	Untreated	38%	62%	13
	Control	67%	33%	49
C	Treated	73%	27%	30
	Untreated	67%	33%	15
	Control	56%	44%	27
D	Treated	56%	44%	66
	Untreated	50%	50%	14
	Control	55%	45%	31
All	Treated	66%	34%	241
	Untreated	51%	49%	76
	Control	64%	36%	148

Table 62

School/Employment Status of Treated, Untreated, and Control Group Members: Second Follow-Up, Third Cohort

Site	Group	% in School or Employed	% Not in School and Unemployed	Sample Size
A	Treated	60%	40%	50
	Untreated	38%	62%	29
	Control	61%	39%	41
B	Treated	61%	39%	89
	Untreated	42%	58%	12
	Control	57%	43%	44
C	Treated	56%	44%	59
	Untreated	63%	37%	27
	Control	53%	48%	38
D	Treated	47%	53%	77
	Untreated	29%	71%	21
	Control	61%	39%	38
All	Treated	56%	44%	275
	Untreated	44%	56%	89
	Control	58%	42%	161

Fourth-cohort school/employment status data are presented for the first (and only) follow-up in Table 63. Both the across-site treated-versus-control and the treatment-versus-control comparisons are statistically significant (Chi Squares = 9.62 and 10.09 respectively, $p < .01$ in both cases). These comparisons are also significant at Site D (Chi Squares = 27.88 and 43.74 respectively, $p < .0001$ in both cases). As was the case with the third cohort, the treatment group at Site D is significantly better off than the treatment groups at the other three sites (Chi Square = 12.21, $p < .001$).

Table 63

School/Employment Status of Treated, Untreated, and Control Group Members: First Follow-Up, Fourth Cohort

Site	Group	% in School or Employed	% Not in School and Unemployed	Sample Size
A	Treated	54%	46%	52
	Untreated	68%	32%	47
	Control	61%	39%	46
B	Treated	61%	39%	51
	Untreated	69%	31%	16
	Control	67%	33%	58
C	Treated	71%	29%	51
	Untreated	50%	50%	8
	Control	55%	45%	22
D	Treated	82%	18%	95
	Untreated	64%	36%	22
	Control	45%	55%	95
All	Treated	69%	31%	249
	Untreated	66%	34%	93
	Control	55%	45%	221

IV. DISCUSSION

The second interim Task B report (Tallmadge & Yuen, 1980) described how implementation events could affect program outcomes. It did not, however, attempt to tie outcome data directly to these events. Such an attempt was made in the present report and a surprisingly high degree of correspondence was found.

In a few instances, outcomes could not be explained in terms of events at the sites. More often, however, they could. Retention rates, for example, were high when the programs were running well and the site climates were positive. They fell with remarkable regularity at times when implementation, staffing, and/or morale problems arose. Similarly, substantial achievement gains in math were observed when qualified math teachers were present. No such gains were observed when math instruction had to be conducted by teachers with other subject-matter specializations.

These relationships between program events and student outcomes are not, and should not be, unexpected. It is eminently sensible that treatment effects should be observed after effective treatments. In the case of the present study, however, these relationships play an unusually important role as one attempts to assess the overall value of the CIP.

There were many implementation problems. They were compounded by unrealistic schedules, uncertain funding, an intrusive evaluation design, and a complex, cumbersome, and somewhat non-responsive decision-making structure. For these reasons, one must consider what might have been, as well as what was, in order to arrive at a fair assessment of the CIP.

All four of the CIP replications experienced periods when the program was being implemented well. Two of the sites had extended periods when, in the opinion of the RMC site visitors, the program was operating in a nearly flawless manner. All four sites also experienced periods of substantial disarray and two of them were "in trouble" during at least half of the demonstration period.

The authors of this report, given the circumstances just described, feel that a fair evaluation of the CIP must consider both the impact of the program when it is being fully implemented and the feasibility of attaining this level of implementation. The latter type of assessment is particularly difficult to make, unfortunately, and depends to a large extent on subjective judgments made by the evaluators.[4] Even if one chooses to ignore considerations of

[4] A full-blown discussion of the feasibility of implementing the CIP is beyond the scope of this report. The final Task A report (Treadway et al., 1981), however, is devoted almost in its entirety to this topic and should be consulted by the interested reader.

implementation feasibility, however, it is important to recognize that outcome measurements taken when a program is not properly or fully implemented reveal little or nothing of what would happen if the same program were implemented as intended.

The results of nearly all of the analyses presented in the preceding chapter were mixed. If, however, one dismisses some of the negative findings as the logical outcomes of poor implementation, the overall picture becomes substantially more positive.

Holding Power

The attrition data are not easy to interpret on an overall basis. Although the treatment and control groups showed approximately equal attrition rates from pre-to-midtest and from pre-to-posttest, the groups were treated differently. Treatment group students who failed to enroll in the CIP or who dropped out or were terminated were systematically excluded from subsequent testings. These individuals were automatically added to the attrition list even though they might have returned for testing had they been allowed to do so (as all members of the control were allowed to do). When this difference is taken into account, it appears that the program does have substantial holding power over its participants.

While the preceding inference is based on somewhat tenuous evidence, it was supported by analyses of individual-site attrition data. There was a remarkably clear pattern of poor implementation being accompanied by high attrition and vice-versa. At least when the programs were functioning well, it seemed that they did a good job of retaining their students

Cognitive Achievement

In the area of reading achievement, the results of the various analyses were somewhat less positive than had been expected. The across-site and across-cohort norm-referenced gains were statistically significant at both mid- and posttest time. At posttest time the gain was also large enough (6.7 NCEs) to be considered educationally significant. The other gain estimates, however, were disappointing. The main question raised by the difference between the norm-referenced and the other gain estimates is which of them is the more credible?

An examination of the data in Tables 3 through 6 reveals that, overall, statistically significant norm-referenced gains were made not only by the treatment group but also by several control and comparison groups. It was the gains made by these other groups that caused the covariance and standardized-gain analyses to produce primarily nonsignificant results, since these approaches yield estimates that are generally quite close in size to the difference between the norm-referenced gains of the treatment group and the

corresponding gains of the control or comparison groups. This same relationship also explains the fact that, where the norm-referenced gain of the comparison group was small (as in the case of the Site A Alternative High School comparison group), the treatment effect estimate derived from the standardized gain analysis was large and statistically significant (see Table 10).

While the relationships between the norm-referenced gain estimates and those produced by the covariance and standardized-gain analyses is understandable, the question remains as to whether the norm-referenced estimates reflect real gains or are the result of some artifact of the study procedure. It must be acknowledged, for example, that the normative interpolations and extrapolations required by the circumstances of the CIP replications, as well as the assignment of students to grade norms based on their ages, may have introduced biases into the norm-referenced evaluation.

If, indeed, the procedures used to implement the norm-referenced analyses introduced bias, then the norm-referenced gain estimates are too high. A more accurate picture of the CIP's impact on reading achievement is then provided by the other analyses. If, on the other hand, one rejects the hypothesis that bias was introduced into the norm-referenced evaluation, one must accept the fact that the gains made by the third-cohort control and comparison groups and by the fourth-cohort control groups were real. This position, in turn, is difficult to accept since there is some doubt that the "treatment" received by most of these groups (Comparison Group 2 at Site B is an exception) was as effective as that of the CIP. In the case of the dropout group in particular, there was presumably no reading-related instruction whatsoever.

One possible explanation is that the gains resulted from operation of the John Henry effect. Another is that there may have been systematic attrition in the control and comparison groups. It does not seem unlikely, in fact, that the members of these groups whose skills had improved would be more highly motivated to attend the posttest session than those who had made no gains. Such students, of course, would be atypical representatives of their groups and would not, therefore, provide a fair baseline against which to measure the impact of the CIP.

It is likely that a related sort of self-selection also occurred in the treatment group. Posttest data, however, were collected from very nearly all of the students enrolled in the CIP at posttest time. These students were, therefore, representative of the group that had received twelve months of the CIP treatment. While they were very likely not representative of the original treatment group, it can be assumed that failure to make substantial gains in reading was probably not a major cause for attrition from the program. For this reason, treatment-group data may be somewhat less biased than control-group data--at least in the area of reading achievement.

Whatever self-selection may have occurred in treatment and control groups during the CIP demonstration most probably resulted from feelings or motivations that were not directly tapped by the instruments used in the evaluation. Nevertheless, a decision was made to explore the possibility that those who dropped out of the groups did differ from those who remained, in terms of the achievement and affective measures that were used. To accomplish this task, mean pretest scores were calculated for two groups: those individuals who were neither mid- or posttested (and thus, presumably had dropped out) and those who had been either mid- or posttested (or both). This was done by site, by cohort and across sites, by cohort using reading, math, internal-external, and self-esteem scores.

There were 20 across-site analyses, two of which were statistically significant at the 5% level. In the third cohort, individuals assigned to the treatment group who did not remain in the program obtained significantly higher math scores than members of the treatment group who did remain. This difference was primarily due to a 9.8 NCE differential observed at Site A. Neither the across-site nor the Site A difference appeared in second- or fourth-cohort data, however.

In the fourth cohort, members of the control group who returned for mid- and/or posttesting had a significantly higher mean score on the Self-Esteem scale than did control students who failed to return. This difference was not present in the third-cohort data. There were also no significant self-esteem differences at individual sites in either the second or third cohorts.

At individual sites, there were four additional statistically significant ($p < .05$) differences. Since 80 comparisons were made, however, 4 is the exact number that would be expected to be "significant at the 5% level" by chance alone.

The attrition analyses, although they did produce a few statistically significant differences, shed little light on possible self-selection biases. While it is interesting that consistent patterns were not found in these analyses, their absence does not remove the possibility that attrition from treatment and control groups was systematic. In fact, the authors believe that at least some of these control group students who returned for mid- and/or posttesting were motivated by competitive feelings, thus producing a John Henry effect.

It is indeed unfortunate that so much speculation is required for the interpretation of the reading (and other) results. The only data, however, that should be free of the various contaminating influences discussed above are those used in the matched-pairs analyses. Unfortunately, there the sample sizes are so small that the gain estimates are necessarily unstable.

The picture was much the same with respect to math. The majority of the norm-referenced and standardized gain analyses showed statistically significant treatment effects--at least at posttest time. On the other hand, only a few of the covariance analyses yielded significant results at midtest time and none was significant at posttest time. The overall, norm-referenced gain estimate at posttest time was 4.3 NCEs, somewhat smaller than the gain in reading but still highly significant. As is pointed out in the Results chapter, the smaller size of the math gain is probably attributable to the difficulty all sites had in hiring and retaining qualified math instructors.

As was the case with reading, several control and comparison groups also made statistically significant, norm-referenced gains. Most frequently in the case of the comparison groups, however, these gains were smaller than those observed in the corresponding treatment groups. As a result, all but two of the comparison group analyses showed statistically significant gains at posttest time.

The gain estimates derived from the matched-pairs analysis were smaller than the others and frequently even negative. All of them, however, were plagued by small sample sizes. The third-cohort, individual-site analyses are illustrative of the kinds of variability that can be expected with such small samples. The apparently large between-site differences are almost certainly meaningless as none of the gain estimates is significantly different from zero.

Career Development Inventory

Most of the analyses performed on the CDI Planning scale showed statistically significant gains both at individual sites and when the data were combined across sites. The situation was slightly less positive for the Resources scale. On the Information scale, the results were generally non-significant at midtest time, but the majority were significant at posttest time.

Care must be taken not to over-interpret the statistically significant gains made by interns on the Planning and Resources scales. The Planning scale in particular does not reflect ability to plan. The scale is made up of such items as, "Talking about my career decisions with an adult who knows something about me." The student response, "I have not given any thought to this" earns one point while the response, "I have done this" earns six points. There are various response options between these two extremes that earn intermediate numbers of points.

It seems to the authors that "gains" on items of this type are more descriptive of the treatment itself than of its impact. It is, for example, an integral part of the CIP for interns to discuss career objectives, plans, and decisions with career developers. It would appear then that any intern who failed to respond, "I have

done this" must have misunderstood the question. Neither the question nor the response, however, gets at the issue of whether the discussion influenced the intern or was useful in any way.

Since the CDI Planning scale contains a significant number of similar items--items that would be expected to show gains simply as a result of participating in the CIP rather than benefiting from it--it must be concluded that the observed gains do not necessarily reflect benefits accrued by the interns.

The CDI Resources scale is made up of similar items and the same argument advanced with respect to the Planning scale is equally applicable. Gains do not necessarily reflect benefits accrued by the interns.

The items that make up the CDI Information scale are of a more traditional nature. They have correct and incorrect response alternatives and tap career-related knowledge. Gains on this scale should, therefore, reflect an actual increase in interns' career awareness.

The study conducted by Gibboney Associates (1977) produced almost identical findings with respect to the CDI. After 10 weeks of program participation, there were significant gains on the Planning and Resources scales and no gain on the Information scale. After a year of program exposure, however, there were small but statistically significant gains on the Information scale. The small size of the gains was explained in terms of mismatch between the career-related instruction provided by the CIP and the questions contained in the test. That argument appears valid--interns learn about specific careers that are of interest to them, while the CDI Information scale is concerned with more general issues such as relationships between aptitudes and types of careers. The failure of the Information scale to show bigger gains should not be interpreted to mean that interns learned little about careers. A more relevant instrument might well have shown much larger gains.

The gains on the Planning and Resources scales should not be dismissed as lightly as the preceding comments might imply. While they reflect changes in exposure rather than the effects of the exposure, the changes are quite large. It is probabaly safe to assume that the exposure had at least some impact, and an optimistic inference might be that the increased exposure contributed significantly to the skills of interns in career planning and in the use of career-related resources.

One final point relating to the Career Development Inventory--the gains on all three scales were uniformly larger at posttest time than they were at midtest time. This pattern, which was also observed in reading and math, suggests that growth proceeds as a direct function of the length of program exposure.

Other Non-Cognitive Measures

Unlike the Gibboney Associates (1977) study, a number of statistically significant gain estimates were found on the Coopersmith Self-Esteem Inventory. More statistically significant gains were found on the Self-Esteem scale in the third-cohort analyses than in either the second- or fourth-cohort analyses. The improved quality (compared to the second cohort) and greater amount (compared to the fourth cohort) of counseling available to third-cohort interns was offered as a possible explanation for this finding. While gains on the Self-Esteem scale amounted to only a few raw-score points and their educational significance may be questionable, the evidence suggests that the influence of the CIP on self-esteem scores was large enough to be reliably measured.

Of some 60 analyses involving the Coopersmith Openness scale, 9 produced statistically significant gains (1 of which favored the control group). Since the goals of the CIP appear unrelated to what this scale measures, no attempt was made to interpret these findings.

With respect to the Rotter Internal/External scale, even fewer of the gains (4 out of 60) were found to be statistically significant. This finding was somewhat surprising since common sense, as well as on-site ethnographic observations (Fetterman, 1981), suggest that long-term participants in the program should feel increased control over the events of their lives. The authors' beliefs on this matter are sufficiently strong, in fact, to lead them to believe that the negative results stem from the fact that the instrument is simply not sensitive to the kinds of changes that occurred.

Follow-Up Outcomes

The follow-up data are more directly related to the stated goals of the CIP than either the attrition or the test score data. One of the program's stated goals is to assist dropouts and potential dropouts to obtain their high school diploma. While the number of actual CIP graduates from the third and fourth cohorts (where control groups were available) was too small to show statistically significant gains, comparisons between treatment and control groups in terms of the number that had graduated from high school, were currently enrolled, or had earned a GED were generally favorable.

For the fourth cohort, the high school status of the treatment group was significantly better than that of the control group at Site D and across sites. This was despite the situation at Site A where the control group presented a better picture than the treatment group (although not significantly so) and significantly better than the control groups at the other three sites ($p < .01$).

The third-cohort data showed a significant advantage for the treatment group over the control group at Site C. The negative results at Site A, however, prevented the difference from being significant overall. When data were combined across the other three sites, a significant advantage was again found for the treatment group.

The second cohort had no control group. A larger percentage of treatment group members had graduated from high school, were currently enrolled, or had earned a GED, however, than was the case with either the third or fourth cohorts. This relationship held at both the first and second follow-ups largely because the results at Site A had not yet turned bad.

The second stated goal of the CIP to which follow-up data were relevant was that of smoothing the transition from school to work. Because large numbers of students were still enrolled in school, however, it seemed most appropriate to compare treatment and control groups in terms of the numbers either in school or employed versus not in school and not employed.

The results of these comparisons were somewhat less favorable than those related to high school status, but still generally encouraging. The fourth-cohort treatment group presented a better picture than the control group both at Site D and overall on the only follow-up that was conducted on that cohort. There were no significant differences between treatment and control groups for the third cohort, but the treated subgroups were superior to the untreated subgroups in both the second and third cohort at the time of the first follow-up.

Perhaps a more positive picture would emerge if we had information regarding the quality of jobs that were held. While queries were made regarding salary levels and probabilities for advancement, too few credible responses were received to show statistically reliable differences between groups.

A Note on Implementing the Evaluation

As pointed out repeatedly throughout the report, this study was plagued by small sample sizes and high (possibly differential) attrition rates. While these conditions seriously restricted RMC's ability to conduct rigorous analyses and to reach conclusions that were unencumbered by excessive numbers of caveats, it is not clear that much could have been done to reduce the problems. Recruiters at all four sites left few, if any, stones unturned in their attempts to attract large numbers of students. In fact, their efforts to meet contractually specified treatment and control group quotas may have been excessively zealous. The authors' impression is that some students were almost literally dragged in and that some of the early attrition stemmed from the fact that these students were never seriously interested in the program.

At mid- and posttest data-collection times, it was possible to test virtually all of the students then enrolled in the CIP. We were unable, however, to obtain high participation rates from students in the control and comparison groups. The students themselves are highly mobile and difficult to track. The resources available for the study were sufficient to support only one half-time and one quarter-time assistant at each site and this manpower level was inadequate for the task. We would recommend at least one full-time and one half-time site assistant at each location.

Another unanticipated problem was that many of the control and comparison group students were enrolled in other schools. Collecting data from them would have been facilitated had we been able to conduct testing sessions in the schools. While some schools were willing to cooperate in this manner, others were reluctant--given that there were no incentives for them to do so. The authors believe that future studies of this type should attempt to arrange an incentive system so that better cooperation can be obtained.

V. SUMMARY AND CONCLUSIONS

The analyses presented earlier in this report provide substantial evidence that the Career Intern Program had a positive impact on participating students. Statistically significant gains were observed on standardized reading and math tests, on all three scales of Super's Career Development Inventory, and in self-esteem. In addition, a significantly larger proportion of the treatment group had graduated from high school, was currently enrolled, or had obtained a GED than was the case for the control group. Evidence with regard to school/employment status was less compelling but still generally positive. Finally, there was evidence that the program was able to retain students--particularly when it was operating well.

The issue of implementation is very important to the understanding and proper interpretation of the study results. When the programs were not functioning smoothly, absenteeism and attrition were high and achievement gains tended to be low. Similarly, when programs had to operate without a qualified math teacher (even if all other aspects of the program were working well and attendance was high) students failed to make significant gains. Gains in self-esteem appeared to require both extended involvement in the program (they did not emerge until posttest time) and extensive contact with qualified counselors.

Relationships of this type were fairly obvious in the data--perhaps because all of the sites experienced substantial implementation problems at various times during the demonstration period. In addition to highlighting relationships, implementation problems also produced negative results. Thus the data should not be taken as an accurate gauge of what the CIP can do. The existing evidence suggests that the program would have had substantially greater impact had fewer implementation problems been encountered.

The initial success at Site B and the delayed but ultimately outstanding performance at Site D stand as testimony that the program can be implemented effectively. The outcome data from those sites at those times are overwhelmingly positive and would seem to provide the best estimate of what the CIP can accomplish.

In addition to problems resulting from incomplete program implementation, the evaluation was hampered by very high attrition rates. At least to some extent, the high attrition resulted from the need to meet contractually specified enrollment quotas that were unrealistic for new and unproven programs. Many students assigned to the treatment group never even enrolled in the program while substantial numbers of others dropped out almost immediately. In any case, one major consequence of the high attrition rate was the threat it posed to the internal validity of the treatment-control evaluation design.

Because of hazards associated with randomized experiments when attrition is high, several other evaluation strategies were also employed. As it turned out, the different strategies yielded somewhat different results. In reading and math, for example, the findings of the norm-referenced evaluations were substantially more positive than those of the covariance and standardized gain analyses which used control and comparison groups respectively. In reading, the norm-referenced gain estimate for the 280 students in the combined third and fourth cohorts was 7.4 NCEs (from the 24th to the 36th percentile) while the corresponding covariance estimate was 2.6 NCEs. In math, the corresponding gains were 4.3 NCEs (from the 12th to the 17th percentile) and 1.4 NCEs, respectively.

The reason for this difference derives from the fact that the control groups also showed positive (norm-referenced) growth in reading and math. It is the authors' opinion that these gain estimates did not arise from biases inherent in the unusual manner in which the norm-referenced evaluation had to be implemented but rather are real. We also believe, however, that the gains did not result from any instructional treatment the control group members received but instead from some combination of a self-selection bias (65% of the control group members chose not to participate in the posttesting session despite a monetary incentive of approximately $20 to do so) and a John Henry effect. The plausibility of the John Henry effect, in turn, derives from the fact that all members of the control group sought, but were denied, admission to the program.

All three scales of the Career Development Inventory showed several statistically significant treatment effects in individual-site analyses. Across sites, the gain estimates were significant in over half of the cases.

Statistically significant gains in self-esteem were observed in half of the covariance and standardized gain analyses at posttest time but in none of the midtest analyses. It was inferred that a substantial amount of treatment is required to effect gains in self-esteem.

Very few of the analyses involving the Rotter Internal-External scale produced statistically significant gains. The authors' own observations, however, and the ethnographic analyses reported by Fetterman (1981) suggest that this finding is misleading. It seems far more likely that CIP students did gain a feeling of control over their lives from the program but that the gain failed to manifest itself in the test scores.

Several of the other instruments used in this study seem less than optimum in retrospect. A particularly salient example is the Information scale of the Career Development Inventory. While statistically significant gains were made on this scale none of them exceeded two raw score points.

All CIP students participate in a semester-long career counseling seminar. In addition, a career-development plan is worked out for each intern. The interns research two career fields in depth and participate in two week-long, Hands-On job experiences. It seems impossible that the total impact of these learning experiences can be reflected by two raw score points. While no more appropriate instrument may be available, it is nearly inconceivable that a better, more relevant one could not be developed. Where the future funding of a program may hinge on the results of an impact evaluation, it seems of utmost importance to employ tests which are relevant to the goals and curriculum of that program.

As regards relevance, it is important to point out here that gains on paper-and-pencil tests such as were used in this study is not a major objective of the CIP. Such gains are, at best, intermediate objectives that may or may not be highly relevant to the program's primary goals of helping participants earn their high school diploma and enhancing their employability. Other data, however, strongly support the CIP's success in achieving the first of these primary objectives and provide at least some support for success in the second.

APPENDIX A

Comparability of the Evaluation Designs Used in the CIP Study

Comparability of Designs

Each of the designs used in this study provides an estimate of the CIP's impact on participating students. That estimate, in turn, is derived from an evaluation of student performance after participating in the program and an estimate of what that performance would have been had the students not participated. The post-participation assessment is the same in all designs, but the "no-treatment expectation" differs.

There is no way of knowing exactly what those students who participated in an experimental treatment would have done had they not participated. It is generally accepted, however, that a good estimate of that performance can be obtained from a similar group of students who did not participate. The credibility of the estimate, of course, depends heavily on the extent to which the two groups are similar.

True Experiments

Randomly assigned groups. One experimental approach that is often used for the purpose of assuring comparability between groups is to randomly assign students drawn from a pool of potential participants to treatment and control groups. Any differences between the groups that result from random assignment can, presumably, be adjusted for through use of covariance analysis.

This so-called classic or "true" experimental design provides unbiased estimates of treatment effects and is generally regarded as preferable to any quasi-experimental design (such as the norm-referenced design). Unfortunately the integrity of the design can be destroyed by attrition. If the students lost from one group are systematically different from those lost from the other, the remaining groups are no longer randomly equivalent, and covariance analysis can no longer adequately adjust for between-group differences.

The matched-pairs design. A variation on the random experiment is one in which pairs of students are formed prior to the assignment process in such a way that their members are as much alike as possible in all ways relevant to the experiment. One member of each pair is then selected randomly for assignment to the treatment group while the other is assigned to the control group.

If the matching is good, initial differences between groups should be close to zero, thus obviating the need for any analysis of covariance-like adjustment. Furthermore, if both members of a pair are discarded when either member is lost through attrition,

the remaining treatment and control groups will still be randomly equivalent. Apart from the practical difficulties associated with implementing this design, its only real drawback is that it is more severely affected by attrition than its less sophisticated counterpart. For example, if attrition results in the loss of 40% of all students, it will result in the loss of 64% of all pairs. The samples remaining for analysis would thus encompass 60% of the original groups for the simple random design but only 36% for the matched-pairs design.

Adjustments for Initial Differences between Groups

Both of the true-experiment designs assess treatment effects through use of a no-treatment expection derived from a sample of students believed to be equivalent to those who actually participated in the treatment. The matched-pairs design does this in a straightforward manner by simply comparing the posttest performance of the two groups. The simple random experiment, on the other hand, may frequently require that an adjustment be made for non-trivial pretest performance differences between groups resulting from the (unmatched) random assignment process.

The assumptions underlying the adjustments that are available to the evaluator may not be met under even the best of conditions. They become increasingly problematical when attrition is high, when there are reasons for suspecting the existence of real differences between the treatment and control groups, or when assignment to the control group may itself affect the behavior of the students.

Under conditions where assignment to treatment and control groups was indeed random and where there was no attrition from either group, analysis of covariance procedures are considered most appropriate to adjust for whatever pre-treatment differences may exist between groups. In the typical two-group (treatment and control) situation, the covariance adjustment entails multiplying the difference between the groups' pretest means by the slope of the common, within-group regression line. (The within-group line is used under the assumption that it is a more accurate and stable estimate of the population value than that provided by either group separately.) The result of this calculation is then used to adjust the posttest means of the two groups.

One major assumption underlies the use of a common, within-group regression line. It is that the two groups are random samples from a single population. If assignment is random, this assumption is, by definition, met at pretest time. The treatment may, however, affect both the mean and the variance of treatment group posttest scores. Under these circumstances, it seems

inappropriate to regard the two groups as random samples from a single population at posttest time. Furthermore, since the slope of the regression line is partially determined by the variance of posttest scores, it becomes seemingly inappropriate to calculate a common, within-group regression line.

If one does not use a combined, within-group regression line to adjust the mean posttest scores of the treatment and control groups, two other particularly interesting possibilities exist. The treatment group's regression line could be used to predict what that group's posttest scores would have been had its pretest score been the same as the control group's. Alternatively, the control group's regression line could be used to predict what that group's posttest score would have been had its pretest score been the same as the treatment group's. Gains would then be calculated by comparing the predicted posttest score of one group with the observed posttest score of the other group.

The gain estimate derived from projected treatment group posttest scores will be different from the one based on projected control group scores unless the two regression lines are exactly parallel. The amount of difference between the two gain estimates will be a joint function of the difference in regression line slopes and the difference in pretest means. In some instances the two gain estimates will differ substantially from one another. Unfortunately, there is no way to determine where "truth" lies. It is perhaps best to regard the two estimates as boundaries defining a range within which the true gain is likely to fall.

All of the covariance analyses included in this report were calculated three different ways: one using a common, within-group regression line; one using the treatment group's regression line in the manner described above; and one using the control group's regression line (also in the manner described above). For treatment-versus-control group comparisons, the tables in the Results section present only the findings of the standard covariance analysis using the common, within-group regression line. However, where the other analyses yielded results that were substantially different, they are discussed in the text.

It was mentioned earlier that an important assumption underlying standard covariance analysis procedures is that the groups being compared be random samples from a single population. Where systematic differences are known to exist between the groups prior to the beginning of the experiment, covariance analysis is thought to systematically underadjust for pretest differences (Campbell & Erlebacher, 1970). Under these circumstances, some form of reliability-corrected covariance analysis (Porter, 1967) or standardized-gain analysis (Kenny, 1975) is generally considered to be more appropriate.

The present study employed standardized-gain analyses in all situations where covariance analysis was also employed. This type of analysis is exactly comparable to covariance analysis except that it makes use of the principal axis of the bivariate distribution of pre- and posttest scores rather than the corresponding regression line. Because three versions of each covariance analysis were carried out, the corresponding three versions of standardized-gain analyses were also conducted (one using the combined, within-group principal axis; one using the treatment group's principal axis; and one using the control group's principal axis).

Considering the covariance and standardized-gain analyses together, six different gain estimates were calculated for each "experiment" (e.g., Site A, treatment group versus regular high school comparison group). The question immediately comes to mind, "Which of the six estimates most accurately reflects the true impact of the program?" If the answer to that question were known, of course, there would be little point in calculating the five less accurate estimates. The answer is not known, however, and therein lies the justification for the multiple analysis approach.

If the smallest of the gain estimates were statistically and educationally significant, one would have a high degree of confidence in labeling the treatment as effective. If five out of the six estimates were not statistically significant, one would have to adopt a more conservative stance. The number of statistically significant estimates thus provides a crude indicator of how much confidence can be placed in the inferences one draws from the analyses. While not "scientific" in any strict sense of the word, considering all six estimates simultaneously is almost certainly a better approach than selecting one as the "best" because the circumstances of this study are such that the assumptions of all of the analyses are violated more often than they are met.

Quasi-Experiments

Because of the high, and probably differential, attrition that occurred between pre- and posttests, it is not entirely clear whether the treatment-control comparisons made in this study should be regarded as true experiments or not. On the other hand, the comparisons made between the treatment groups and the specially selected comparison groups at each site cannot be regarded as true experiments. They are best categorized as a class of quasi-experiments called the non-equivalent control group design.

The non-equivalent control group design. As pointed out above, the comparison (as opposed to control) groups used in this

study cannot be considered random samples from the same population from which the treatment group was drawn. It is to be expected that they differ from the treatment group in systematic ways and are samples from different populations. For this reason, the standardized-gain approach was considered preferable to the covariance-analysis approach as a strategy to adjust for pretest differences between groups. The treatment-versus-comparison group analyses in the tabular presentations of the Results section of this report thus reflect that mode of analysis.

As was the case with the treatment-versus-control comparisons (where analysis of covariance results are presented in the tables), however, analyses were conducted using all six of the adjustment strategies described earlier in this appendix. Where results from the other analyses differed substantially from the standardized-gain results, the differences are discussed in the text.

It should be pointed out that quasi-experiments attempt to provide answers to questions that are somewhat different from those addressed by true experiments. The latter generate estimates of what the treatment group's performance would have been in the absence of the treatment. Quasi-experiments simply compare the posttest performances of the treatment group with that of another, similar group. They either assume that the groups were equal in pre-treatment performance levels or they statistically adjust post-treatment measures to compensate for pretest differences.

The assumption is often made that the posttest (or adjusted posttest) performance of the comparison group provides a good approximation of a no-treatment expectation for the treatment group. It would be more prudent, however, to acknowledge that quasi-experiments really address the question, "How much better (or worse) would the treatment group have performed than the comparison group if the two groups had started out equal?" If that orientation is taken, the obtained results can be interpreted in terms of the similarities and differences between the groups and additional insights may be obtained.

The norm-referenced design. The norm-referenced design assesses treatment effects in terms of changes in status with respect to the national norms from pre- to posttest. If a group's mean pretest score placed it at the 20th percentile prior to participation in the program being evaluated and its mean posttest score placed it at the 25th percentile, the 5-percentile gain would be attributed to the effect of the treatment. In essence, the design compares the growth of treatment-group students with students at the same pretest achievement level attending a nationally representative sample of schools.

The design does not normally provide a local no-treatment expectation since treatment-group students, if they did not participate in the treatment, would not be attending a nationally representative sample of schools. While, from some perspectives, this characteristic of the norm-referenced design might be viewed as an advantage, it does make the design systematically different from designs that use local control or comparison groups.

The evaluation findings presented in the Results section of this report show several instances where substantial differences exist between the norm-referenced gains and the gains derived from control or comparison group analyses. In these cases, it is interesting to examine the norm-referenced gains made by the control or comparison group (these gains are also included in the tables).

Subtracting the norm-referenced gain made by the control or comparison group from the norm-referenced gain made by the treatment group yields a treatment-effect estimate that very closely approximates the estimate derived from the corresponding covariance or standardized-gain analysis. When used in this manner, the norm-referenced model does provide a local no-treatment expectation. The feature that is the primary contributor to the design's desirability, however, is its ability to produce a gain estimate without requiring a control or comparison group. Under these circumstances, of course, it does not provide a local no-treatment expectation.

APPENDIX B

Selection of the Achievement Test to be Used in the CIP Evaluation Study

SELECTION OF THE ACHIEVEMENT TEST TO BE USED IN THE CIP EVALUATION STUDY

The test used to evaluate the achievement gains produced by the CIP should possess several important characteristics. To conduct a norm-referenced evaluation the test must have empirical normative data at grades nine, ten, eleven, and twelve, based on nationally representative samples of students. To be sensitive to project impact, the content of the tests should not be uninteresting, esoteric, or irrelevant to the students in CIP. It should reflect as closely as possible the emphasis of the CIP instruction. The level of test selected should be appropriate for the functional level of the students. The test should not be so difficult that the average score of the group tested is at chance nor should it be so easy that, on the average, students answer more than 75% of the items correctly. It would also be desirable for the test to have empirical normative data at more than one point during the year. The number of test items and time required to take the test should fall within reasonable limits and the format of the test booklets should be attractive and easy to follow.

In the review process the following tests were examined: California Achievement Test (1970 and 1977), Comprehensive Tests of Basic Skills (1968 and 1973), Diagnostic Mathematics Inventory (1975), Gates-MacGinitie Reading Tests (1964), Iowa Tests of Basic Skills (1971), Metropolitan Achievement Tests (1970 and 1978), Prescriptive Reading Inventory (1975), Sequential Test of Educational Progress (1969), SRA Achievement Series (1971), and Stanford Achievement Tests (1973).

Of this group, only five tests were found to have normative data at grades nine, ten, eleven, and twelve. Specifically, the California Achievement Tests (1970 and 1977), Comprehensive Tests of Basic Skills (1973), Metropolitan Achievement Tests (1978) and the Sequential Tests of Educational Progress (1969) fulfilled this requirement.

Each of the five tests was examined in detail. The times of the year when the test was normed and the forms that are available were noted. The level of the test intended for high school students and the next lower (or easier) level of the test was determined. For each level, the number of items in each subtest, the time required to take the test, and the length and topic of each passage were listed. A summary of this information is provided for each test (see Figures 1 through 5).

This review revealed some significant differences among the five tests. The passages in the STEP II subtest are longer

Level	Empirical Norming Dates	Forms
4	6.7, 7.4, 7.7, 8.4, 8.7, 9.4, 9.7	A & B
5	9.7, 10.4, 10.7, 11.4, 11.7, 12.4, 12.7	A & B

Level 4

	Reading Vocab.	Comp.	Math Comp.	Concepts & Problems
No. of Items	40	45	48	50
Testing Time (min)	10	40	28	23

Level 5

	Reading Vocab.	Comp.	Math Comp.	Concepts & Problems
No. of Items	40	45	48	50
Testing Time (min)	10	40	33	22

Content of Level 4 - Reading Subtest

Vocab.
2- or 3-word phrases, find synonym for work in boldface

Reading Comp.
Example of Table of Contents
Example of Index
5 paragraphs - composition of planet earth, volcanoes, earthquakes
7 paragraphs - passage about the need to conserve resources
4 paragraphs - the laser--its history and use
2 paragraphs - logic statements--diagram of a "statement of order"

Content of Level 5 - Reading Subtest

Vocab.
2- or 3-word phrases, find synonym for word in boldface

Reading Comp.
Questions about using a book--glossary, appendix, bibliography
5 paragraphs - the scientific method vs. authoritarianism
8 long paragraphs - Bill of Rights
4 paragraphs - studying the ocean floor
4 paragraphs - aptitude measures--kinds, use of results
7 paragraphs - logic statements--i* = i normal; i* = i abnormal then ...

Figure 1. Summary of content and other characteristics of the California Achievement Test (1970)

Level	Empirical Norming Dates	Forms
18	7.7, 8.1, 8.7, 9.1, 9.7, 10.1	C & D
19	9.7, 10.1, 10.7, 11.1, 11.7, 12.1, 12.7	C & D

Level 18

	Reading Vocab.	Comp.	Math Comp.	Concepts & Problems
No. of Items	30	40	40	45
Testing Time (min)	10	35	25	35

Level 19

	Reading Vocab.	Comp.	Math Comp.	Concepts & Problems
No. of Items	30	40	40	45
Testing Time (min)	10	35	25	35

Content of Level 18 - Reading Subtest

Vocabulary
2- or 3-word phrases are presented. Student is to find synonym of underlined word in phrase

Reading Comp.
5 paragraphs - the story of Maria Mitchell, the astronomer (has a picture)
1 paragraph - radio commercial advertising Valley Music Store
2 paragraphs - salesman's speech offering a $3.00 surprise
4 stanzas - poem about storms
4 paragraphs - history of guitar (pic. of instruments preceding the guitar)
3 paragraphs - newspaper article about proposed route for state highway and letters written in response--1 pro, 1 con
4 paragraphs - captain's log describing trip to rescue survivors

Content of Level 19 - Reading Subtest

Vocabulary
Same as Level 18

Reading Comp.
7 paragraphs - report of a dream--dreamed in a sleep and dream lab (has fantasy picture)
3 paragraphs - editorial about importance of eating natural foods
3 paragraphs - speech given by high school student about contributing to student community garage (pic. of student addressing group)
5 paragraphs - description of sun, solar energy, and sun's rays
3 stanzas - about skyscrapers
6 long paragraphs - work and life of Orozco the artist
1 paragraph - radio ad about Tuff Tape

Figure 2. Summary of content and other characteristics of the California Achievement Test (1977)

Level	Empirical Norming Dates	Forms
3	6.7, 7.7, 8.7	S & T
4	8.7, 9.7, 10.7, 11.7, 12.7	S & T

Level 3

	Reading Vocab.	Comp.	Math Comp.	Concepts & Problems
No. of Items	40	45	48	50
Testing Time (min)	12	35	40	35

Level 4

	Reading Vocab.	Comp.	Math Comp.	Concepts & Problems
No. of Items	40	45	48	50
Testing Time (min)	11	35	40	30

Content of Level 3 - Reading Subtest

Vocabulary
Find synonym

Reading Comp.
5 paragraphs - girl willing to keep her promise to babysit even though she would rather go to the rock festival
1 paragraph - very difficult paragraph about ability to tell history of an abandoned farm by studying landscape
2 paragraphs - about the effects of a meteorite crashing to earth in 1947
Swimming pool rules--questions deal with results of breaking rules
5 paragraphs - story about efforts of junior high school students to make community aware of pollution, etc. through "earth day"
5 paragraphs - about change in English language from time of Old English
3 stanzas - a poem about autumn

Content of Level 4 - Reading Subtest

Vocabulary
Select synonyms

Reading Comp.
5 paragraphs - thoughts of swimmer before he swims his race--200 butterfly
3 paragraphs - discusses the idea of "humanness" in animals and objects
5 long paragraphs - Margaret Mead's study of Samoan culture--ways in which individuals learn values from group
6 paragraphs - communes--reason for development and their advantages
Poem - expressing sympathy with caged birds
5 paragraphs - the many choices offered to high school graduates in terms of further education
3 paragraphs - shrews hunting for food

Figure 3. Summary of content and other characteristics of Comprehensive Tests of Basic Skills (1973)

Level	Empirical Norming Dates	Forms
Adv. 1	7.1, 7.7, 8.1, 8.7, 9.1, 9.7	JS KS
Adv. 2	10.1, 10.7, 11.1, 11.7, 12.1, 12.7	JS KS

	Level Adv. 1		Level Adv. 2	
	Reading Comp.	Math	Reading Comp.	Math
No. of Items	55	50	50	50
Testing Time (min)	35	40	30	40

Content of Adv. 1 - Reading Subtest

Reading Comp.

1 paragraph - passage about marmalade
2 paragraphs - passage about skin diving
4 paragraphs - passage about using the city streets as a playground and the benefits of sports activities
2 paragraphs - very simple summary of Shakespeare's Pyramus and Thisbe
4 paragraphs - formation of Sherlock Holmes clubs
4 paragraphs - girl's reaction to receiving a Christmas gift that is a great disappointment
2 paragraphs - invention of yoyo
1 paragraph - Mary Shelley's writing of Frankenstein
3 paragraphs - Leonardo de Vinci--life and work

Content of Adv. 2 - Reading Subtest

Reading Comp.

2 paragraphs - "Babe" Zaharias, the athlete
3 paragraphs - use and history of passwords to identify friends vs. foes--"shibboleth"
3 paragraphs - development of Monopoly game
1 paragraph - discus throwing--includes many numbers about size, distance, etc.
3 paragraphs - "familiar strangers", definition, results of psychological study of commuters
1 paragraph - effects of wind and water on earth and trees
1 paragraph - unpopular boy who is a bookworm
1 paragraph - description of hostels

Figure 4. Summary of content and other characteristics of Metropolitan Achievement Test (1978)

Level	Empirical Norming Dates	Forms
2	9.7, 10.7, 11.7, 12.7	A & B
3	6.7, 7.7, 8.7	

	Level 2		Level 3	
	Vocabulary	Reading Comp	Vocabulary	Reading Comp
No. of Items	30	30	30	30
Testing Time (min)	15	30	15	30

Content of Level 2 - Reading Subtest

Vocab.
Two types of items: Sentences presented and second sentence must be completed using one of four choices. Word used in sentence and find its synonym

Reading Comp.
5 paragraphs - describes life of chickadees
9 paragraphs - from Charles Dickens' Bleak House--has old-fashioned dialogue
8 stanzas - dog and man are friend, have fight, dog bites man, but dog dies
3 long paragraphs - groups in past may be thought more noble than they were viewed by their contemporaries (e.g., "knights")
4 long paragraphs - use of symbols
Dialog from a play - idiosyncracies of a will that must be fulfilled in order to inherit the money

Content of Level 3 - Reading Subtest

Vocab.
Same as Level 2

Reading Comp.
3 paragraphs - discovery and use of glass to magnify objects
9 paragraphs - the story of Orpheus from greek mythology--the importance of music
5 paragraphs - the composition of glass, glassblowing
6 paragraphs - history of Vietnamese people
7 stanzas - poem about forgetting
7 paragraphs - about kidnapping of young Gilbert who later becomes composer of Gilbert & Sullivan fame

Figure 5. Summary of content and other characteristics of Sequential Tests of Educational Progress (1969)

than the passages of the others, and the content appears more difficult. The STEP II norms are based on the performance of students who were tested almost ten years ago. Using "old" norms may produce misleading achievement status information in norm-referenced evaluations. In addition, empirical data are provided for only one time of the year. Of the five tests, the STEP II appeared to be the least desirable.

A drawback of the CAT '70 is that reading passages of both levels include questions about using parts of books (table of contents, index, etc.) to find information. These questions would seem to be more appropriate in a subtest covering reference skills rather than reading comprehension. In addition, the reading subtests present diagrams of logical relationships from which the students are asked to draw logical conclusions. This may be a foreign task to many students. Finally, since there is a more recent edition of the CAT it would be preferable to use the 1977 edition instead of the 1970. For these reasons, the CAT 70 was felt not to be the best test to use for the evaluation.

For the CTBS '73, the passages in Level 3 (the level we would most likely use) are ordered so that two of the more difficult ones are presented first. This order of presentation may discourage students so that either they will not respond to the remaining items or they may respond at random. A second drawback of the CTBS '73 is that empirical normative data are available for only one month of the year.

The MAT '78 and CAT '77 are the newest of the achievement tests reviewed. Both tests have empirical normative data for October and April. A cursory examination of the content of the reading tests of both the MAT '78 and the CAT '77 showed that either one would be appropriate to use in the CIP evaluation. The passages in the CAT '77, however, seem to be more relevant and inherently more interesting than those of the MAT '78. For example, the radio advertisement passage, the salesman's speech, and the newspaper editorial all present material that reflects "real world" situations that students are likely to have encountered. Of course, it also has passages that are probably of less interest -- the story of a woman astronomer, the history of the guitar, and a poem about storms. The majority of the passages in the MAT '78 deal with topics that would not be of concern to CIP interns. For example, there are passages about marmalade, skin diving, and Leonardo da Vinci.

At a more detailed level the two tests were studied in terms of the instructional objectives that each test attempts to measure. In each test's manual, the instructional objectives

upon which the test was constructed are listed and the test items that measure each objective are identified. These are presented below in Tables 64 and 65. Although the objectives selected by the two publishers do not match perfectly, by collapsing some sub-objectives and relabeling others, it is possible to make comparisons between the tests. (It should be noted that the MAT '78 does not offer a separate vocabulary subtest. Vocabulary items are included in the reading comprehension section.) Direct comparisons can be made between the two tests as to the number of vocabulary items each contains and the number of items asking for literal information. After examining the test items, the MAT inferential category of objectives appears to be equivalent to the CAT interpretive category, and the MAT evaluative category appears to be equivalent to the CAT critical category.

The number and percentage of items under each objective are presented by test in Table 66. The greatest difference in content between the two tests is in the number of items covering literal meaning. The MAT has over three times as many items as the CAT. A second difference between the tests is that the CAT has over twice as many critical thinking items as the MAT. Assuming that CIP reading instruction focuses more on teaching students to grasp the literal meaning rather than the implications of what they read, this analysis indicates that the MAT would be the more appropriate test to give.

A similar type of comparison was made between the Mathematics subtests of the CAT '77 and the MAT '78, as shown in Tables 67 and 68. The CAT offers two separate subtests: Mathematics Computations and Mathematics Concepts and Applications. The MAT has placed both types of items in a single subtest. Concepts and applications problems are the first 32 items and computation problems are the remaining 18.

The two tests are similar in all areas except the number of computation problems involving fractions and decimals, geometry and measurement, and numeration. The difference can be attributed to the fact that the CAT has 35 more items than the MAT, and they are distributed over these three objectives. Although the MAT is a shorter test; it is claimed by its publishers to be as reliable as the other major achievement tests.

Conclusions

Either the CAT '77 or the MAT '78 would be suitable for use in the evaluation of the CIP. Only one test can be selected. After detailed review of both tests, the MAT '78 was chosen over the CAT '77. The reasons for this decision are summarized below.

Table 64

MAT '78 Advanced Level 1, Form JS, Reading Comprehension Test Items Grouped by Instructional Objective and by Passage

		Literal		Inferential		
Passage	Vocab.	Specific	General	Specific	General	Evaluative
1	6	1	5	2,3	4	
2		7,8,10	9	12	11	
3	18	14,15,17		13		16
4		19	21,23	20,22		24
5	30	25,27	26	28,29		
6		32,34	31	33,35	36	
7	41,42		39	38		37,40
8	48	43,44,46		45		47
9		49,50,52		51,55	53,54	

Table 65

CAT '77 Level 18, Form C, Reading Comprehension Test Items Grouped by Instructional Objective and by Passage

	Vocab.	Literal		Interpretive		Critical	
	Syn. Ant.	Recall	Inferred	Character	Figurative	Author	Per-
Passage	Multi	of facts	Meaning	Analysis	Language	Att.	suasion
1		31,36	34	32,33,35,37			
2							38-40
3							41-433
4					44-50		
5		51,52,54,56	53,55,57				
6						58-63	
7		64	65,67,69	66,68,70			
0	1-20,21-25, 26-30						

Table 66

Number and Percentage of Items Under Each Objective

	MAT '78		CAT '77	
Objective	N	%	N	%
Vocabulary	6	11	30	43
Literal	25	45	7	10
Inferential/Interpretive	19	35	21	30
Evaluative/Critical	5	9	12	17
Total	55	100	70	100

Note: CAT '77 has a total of 70 items, including separate subtests for vocabulary and reading comprehension.
MAT '78 has a total of 55 items, vocabulary and reading comprehension items are together in a single subtest.

Table 67

MAT '78--Advanced Level 1, Form JS, Mathematics
Item Number and Number of Items Under Each Objective

Objective	Item Number	Number of Items Measuring Objective
Graphs & Statistics	30,31,32,25,26,27	6
Fractions & Decimals	41-50	10
Laws & Properties	15-18	4
Whole Numbers	33-40	8
Problem Solving	1-6	6
Geometry & Measurement	19-24, 28, 29	8
Numeration	7-14	8

Table 68

CAT '77--Level 18, Form C, Mathematics Computations and Mathematics Concepts and Applications
Item Number and Number of Items Under Each Objective

Objective	Item Number	Number of Items Measuring Objective
Graphs & Statistics (Functions & Graphs)	55,59,66,83	4
Fractions & Decimals (Math Computation)	1,2,4,8,9,10,14,15,19,20, 21-24,26,27,29,30,31-40	28
Laws & Properties (Math Computation)	13,18,25,28	4
Whole Numbers (Math Computation)	3,5,6,7,11,12,16,17	8
Problem Solving (Story Problems)	53,65,70,75,76,77,78	7
Geometry & Measurement	45,46,48-50,58,60,72,73, 78-80,82-84	16
Numeration	41-44,47,51,52,56,57,62-64,	18

NOTE: The objectives in parentheses are the labels used by the publisher of CAT '77.

The primary disadvantage of the Metropolitan is that its content appears less interesting than that of the CAT '77 and as a result of this, interns may not be as motivated to take and complete the test. However, the test items of the CAT '77 include a greater number of higher-level thinking questions than the MAT '78. Compared to the MAT, the California has a much larger proportion of test items that require the reader to make an evaluation or critical interpretation of a passage. The Metropolitan Achievement Test, in contrast to the California, has a much larger proportion of test items that require the reader to make a literal interpretation. Whereas the CAT passages may be more entertaining to read than the MAT's, the test questions are more difficult.

A second difference between the two tests is the way in which the test items are ordered. The questions about any one passage of the CAT are likely to come from one category of instructional objective. For example, in the CAT all of the questions about passage 3 concern critical thinking and all those about passage 4 concern figurative language. In the MAT, test questions or a single passage always cover more than one instructional objective. For example, the questions for passage 3 cover vocabulary and literal, inferential, and evaluative thinking. A student taking the CAT who finds it difficult to respond to questions that require critical thinking may miss all the items about one passage and may become discouraged about attempting more items. If the same student were to take the MAT and were to incorrectly answer similar types of items, the errors will be scattered throughout the test. The arrangement of the MAT test items would seem superior to that of the CAT.

An additional advantage of the MAT that has not been emphasized is that it requires less time to administer. The MAT reading subtest takes 35 minutes compared to 45 for the CAT; the MAT mathematics subtest requires 40 minutes versus 60 minutes for the CAT.

The MAT also fulfills the other criteria that were listed at the beginning of the paper. It has empirical norms for October and April for grades 9, 10, 11, and 12. It is constructed so that the level of test that is appropriate to the functional level of the students can be administered and it is still possible to compare their test performance to that of grade-level peers.

APPENDIX C

Instruments

CAREER DEVELOPMENT INVENTORY

FORM I

RMC Research Corp.
2570 W. El Camino Real
Mountain View, CA 94040
415/941-9550

CAREER DEVELOPMENT INVENTORY

FORM I

DONALD E. SUPER, ET AL.

TEACHERS COLLEGE, COLUMBIA UNIVERSITY
NEW YORK, NEW YORK

* * * * * * * * * * * * * * *

INTRODUCTION

The questions you are about to read ask you about school, work, your future career, and some of the plans you may have made. The only right answers are the ones which are right for you. Later, some questions ask about career facts; others ask you to judge students' plans. Give the best answers you can.

Answers to questions like these can help teachers and counselors offer the kind of help which high school students want and need in planning and preparing for a job after graduation, for vocational and technical school training or for going to college.

ANSWER ALL QUESTIONS. If you are not sure about an answer, guess. There is no time limit, but work as rapidly as you can; the first answer that comes to you is often the best one.

NAME_____GRADE_____DATE_____

YOUR FUTURE OCCUPATION

In your present thoughts and plans, what kind of work would you like to do when you finish all of your education and training? What kind of occupation do you plan to enter? (For example: bookkeeper, machinist, lawyer, registered nurse, small store owner, waitress, engineer, shop foreman, elementary teacher, truckdriver, etc.) Write the name(s) of the occupation(s) you have thought about on the lines below.

1st choice_____

2nd choice_____

3rd choice_____

4th choice_____

The questions begin on the next page. Mark them according to the instructions at the top of each section.

I. How much thinking and planning have you done about your educational and occupational future? What kinds of plans do you have? For each of the 14 statements below, choose one of the following 6 answers to show what you have done about what is mentioned in the statements. Place the number of your answer in the space to the left of each statement.

Here are the possible answers:

1 -I have not given any thought to this.

2 -I have given some thought to this, but haven't made any plans to do this.

3 -I have some plans to do this, but am still not sure of them.

4 -I have made definite plans to do this, but don't know how to carry them out.

5 -I have made definite plans to do this, and know what to do to carry them out.

6 -I have done this.

Here are the statements:

____ 1. Finding out about different kinds of educational and occupational possibilities by going to the library, sending away for information concerning the different possibilities, or talking to somebody who knows about the possibilities.

____ 2. Talking about my career decisions with an adult who knows something about me.

____ 3. Taking courses which will help me decide what line of work to go into when I leave school or college.

____ 4. Taking courses which will help me in college, in job training, or on the job.

____ 5. Taking part in school or out-of-school activities which will help me in college, in training, or on the job.

____ 6. Taking part in school or after-school activities (for example: science club, school newspaper, Sunday School teaching, volunteer nurse's aide) which will help me decide what kind of work to go into when I leave school.

____ 7. Getting a part-time or summer job which will help me decide what kind of work I might go into.

____ 8. Getting a part-time summer job which will help me get the kind of job or training I want.

Here are the possible answers:

1-I have not given any thought to this.

2-I have given some thought to this, but haven't made any plans to do this.

3-I have some plans to do this, but am still not sure of them.

4-I have made definite plans to do this, but don't know how to carry them out.

5-I have made definite plans to do this, and know what to do to carry them out.

6-I have done this.

Here are some more statements:

____ 9. Getting money for college or training.

____10. Dealing with things which might make it hard for me to get the kind of training or the kind of work I would like.

____11. Getting the kind of training, education, or experience which I will need to get into the kind of work I want.

____12. Getting a job once I've finished my education and training.

____13. Doing the things I need to do to become a valued employee who doesn't have to be afraid of losing his job or being laid off when times are hard.

____14. Getting ahead (more money, promotions, etc.) in the kind of work I choose.

15. How would you rate your plans for "after high school"? (Please check (✓) one answer.)

a. _____ Not at all clear or sure

b. _____ Not very clear

c. _____ Some not clear, some clear

d. _____ Fairly clear

e. _____ Very clear, all decided

II. Students differ greatly in the amount of time and thought they give to making choices. Use the five ratings below to compare yourself to the typical students of your sex in your grade in each of the areas of choice listed below. Mark the number of your rating in the space provided in each statement.

Here are the ratings:

1 - much below average, not as good as most

2 - a little below average

3 - average

4 - a little above average

5 - much above average, better than most

Here are the statements:

16. Compared to my classmates I am ____ in the amount of time and thought I give to choosing high school courses.

17. Compared to my classmates I am ____ in the amount of time and thought I give to choosing high school activities.

18. Compared to my classmates I am ____ in the amount of time and thought I give to choosing out-of-school activities.

19. Compared to my classmates I am ____ in the amount of time and thought I give to choosing among general alternatives available to me after high school (for example: choosing college or business school or technical school or work or military service or marriage, etc.)

20. Compared to my classmates I am ____ in the amount of time and thought I give to choosing among specific alternatives available to me (for example: type of college, branch of the military service, characteristics of husband or wife, etc.)

21. Compared to my classmates I am ____ in the amount of time and thought I give to choosing an occupation for after high school, college or job training.

22. Compared to my classmates I am ____ in the amount of time and thought I give to choosing a career in general.

III. How much do you know about the occupation you said you would most like to enter on page one of this inventory. Below are five possible answers to use in answering statements 23 though 33. Mark the number of your answer in the space provided in each statement.

Here are the answers:

1 - hardly anything

2 - a little

3 - an average amount

4 - a good deal

5 - a great deal

Here are the statements:

23. I know ____ about what people really do on the job I said I would like to enter.

24. I know ____ about specialities in the occupation I said I would like to enter.

25. I know ____ about different places where people might work in this occupation.

26. I know ____ about the qualifications and skills needed for this occupation.

27. I know ____ about the environmental working conditions in this occupation.

28. I know ____ about the education or training needed to get into this occupation.

29. I know ____ about the courses offered in high school that are the best for this occupation.

30. I know ____ about the need for more people in this occupation

31. I know ____ about different ways of getting into this occupation.

32. I know ____ about the starting pay in this occupation.

33. I know ____ about the chances for getting raises and promotions.

IV. What sources of information would you go to for help in making your job or college plans? Use the five possible answers listed below to show whether or not you would go to the sources of information listed below. Mark the number of your answer in the space provided in each statement.

Here are the answers:

1 - definitely not

2 - probably not

3 - not be sure whether to

4 - probably

5 - definitely

Here are the statements:

34. I would ____ go to my father or male guardian.

35. I would _____ go to my mother or female guardian.

36. I would _____ go to my brothers, sisters, or other relatives.

37. I would _____ go to my friends.

38. I would _____ go to coaches of teams I have been on.

39. I would _____ go to my minister, priest, or rabbi.

40. I would _____ go to teachers

41. I would _____ go to school counselors.

42. I would _____ go to private counselors, outside of school.

43. I would _____ go to books with the information I need.

44. I would _____ go to audio or visual aids like tape recordings, movies or computers.

45. I would _____ go to college catalogues.

46. I would _____ go to persons in the occupation or at the college I am considering.

47. I would _____ go to TV shows, movies, or magazines.

V. Here again are five answers which are to be used with statements 48 through 61. This time use the answers to show which of the sources of information below have already given you information which has been helpful to you in making your job or college plans. Mark the number of your answer in the space provided in each statement.

Here are the answers:

1 - no useful information

2 - very little useful information

3 - some useful information

4 - a good deal of useful information

5 - a great deal of useful information

Here are the statements:

48. I have gotten _____ from my father or male guardian.

49. I have gotten _____ from my mother or female guardian.

50. I have gotten _____ from my brothers, sisters or other relatives.

51. I have gotten _____ from my friends.

52. I have gotten _____ from coaches of teams I have been on.

53. I have gotten _____ from my minister, priest, or rabbi.

54. I have gotten _____ from teachers.

55. I have gotten _____ from school counselors.

56. I have gotten _____ from private counselors, outside of school.

57. I have gotten _____ from books with the information I needed.

58. I have gotten _____ from audio or visual aids like tapes recordings, movies, or computers.

59. I have gotten _____ from college catalogues

60. I have gotten _____ from persons in the occupation or at the college I am considering.

61. I have gotten _____ from TV shows, movies, or magazines.

VI. Here each question has its own set of possible answers. Check (✓) only one answer for each question.

62. Which one of the following is the best source of information about job duties and opportunities?

____ 1) The Encyclopedia Britannica

____ 2) World Almanac

____ 3) Scholastic Magazine

____ 4) The Occupational Index

____ 5) The Occupational Outlook Handbook

63. Which one of the following would be most useful for detailed information about getting into college?

____ 1) The World Book Encyclopedia

____ 2) Webster's Collegiate Dictionary

____ 3) Lovejoy's College Guide

____ 4) Reader's Digest

____ 5) The Education Index

64. Which one of the following pairs of occupations involves the same level of training and responsibility?

_____ 1) Tailor, Sales Clerk

_____ 2) Engineer, Banker

_____ 3) Tailor, Engineer

_____ 4) Banker, Sales Clerk

65. The occupational fields expected to grow most rapidly during the next ten years are:

_____ 1) Professional and service

_____ 2) Sales and crafts

_____ 3) Crafts and clerical

_____ 4) Labor and sales

66. Between 1910 and 1970, the industry employing the greatest number of workers changed from:

____ 1) Argriculture to wholesale and retail trade

____ 2) Manufacturing to agriculture

____ 3) Wholesale and retail trade to manufacturing

____ 4) Agriculture to manufacturing.

VII. Occupations differ in the amount and type of education required for employment. Select the type of education required for each of the occupations below and mark the number of your answer in space to the left of each statement.

Type of Education:

1 - High School Graduation

2 - Apprenticeship Training

3 - Technical School or Community College (2 year)

4 - College Degree (4 year)

5 - Professional Degree Beyond College

Occupations:

____ 67. Stenographer

____ 68. Dental Technician

____ 69. Family Doctor (Physician)

____ 70. Mail Carrier

____ 71. Plumber

____ 72. Computer Operator

____ 73. Bank Clerk

____ 74. Social Worker

VIII. Many occupations use special tools. Below is a list of occupations and a list of special tools or equipment. Match the occupation with its equipment by marking the number of the appropriate equipment in the space to the left of the occupation.

Type of Equipment:

1 - Manikin

2 - Ammeter

3 - Centrifuge

4 - Trowel

5 - Ledger

Type of Occupations:

____ 75. Electrician

____ 76. Bookkeeper

____ 77. Bricklayer

____ 78. Dressmaker

____ 79. Medical Technician

IX. Here again, each question has its own set of answers. Check (✓) only one answer for each question.

80. In the 9th and 10th grades, plans about jobs and occupations should:

____ 1) be clear.

____ 2) not rule out any possibilities.

____ 3) keep open the best possibilities.

____ 4) not be something to think about.

81. Decisions about high school courses can have an effect on:

____ 1) the kind of diploma one gets.

____ 2) the kind of training or education one can get after high school.

____ 3) later occupation choices.

____ 4) how much one likes school.

____ 5) all of these.

82. Decisions about jobs should take into account:

____ 1) strengths, or what one is good at learning and doing.

____ 2) what one likes to do.

____ 3) the kind of person one is.

____ 4) the chances for getting ahead in that kind of job.

____ 5) all of these.

83. One of the things that great artists, musicians, and professional athletes have in common is the desire to:

____ 1) make money.

____ 2) have large audiences.

____ 3) be the best there is at what they do.

____ 4) teach others what they do.

84. Mary thinks she might like to become a computer programmer, but she knows little about computer programming. She is going to the library to find out more about it. The most important thing for Mary is know now is:

____ 1) what the work is, what she would do in it.

____ 2) what the pay is.

____ 3) what the hours of work are.

____ 4) where she can get the right training.

85. Jane likes her high school biology and general science courses best. She likes to do her schoolwork alone so she can concentrate. When she begins to think about her future occupation, she should consider:

____ 1) Nurse.

____ 2) Accountant.

____ 3) Medical Laboratory Technician.

____ 4) Elementary School Teacher.

86. Peter is the best speaker on the school debating team. The school yearbook describes him as "our golden tongued orator--a real nice guy who can listen as well as talk--he could sell refrigerators to the Eskimos." Peter will probably graduate in the bottom half of his class, although his test scores show that he is very bright. His only good grades (mostly B's) are in business subjects. His poorest grades are in English and social studies (mostly C's).

Peter's desire to become a trial lawyer is not very realistic because:

____ 1) with his grades he will have difficulty getting into a four-year liberal arts college.

____ 2) he has poor grades in the subjects that are most important for law.

____ 3) there is much more to being a lawyer than being good at public speaking.

____ 4) all of the above are good reasons for thinking that Peter will have a hard time becoming a trial lawyer.

87. The facts about Peter suggest that he should think about becoming:

____ 1) an accountant.

____ 2) a salesman.

____ 3) an actor.

____ 4) a school counselor.

____ 5) a lawyer.

88. Ernie took some tests which show that he might be good at clerical work. Ernie says, "I just can't see myself sitting behind a desk for the rest of my life. I'm the kind of guy who likes variety. I think being a traveling salesman would suit me fine." He should:

___ 1) disregard the tests and do what he wants to do.

___ 2) do what the tests say since they know better than he does what he would be good at.

___ 3) look for a job which will let him use his clerical abilities but not keep him pinned to a desk.

___ 4) ask to be tested with another test since the results of the the first one are probably wrong.

89. Joe is very good with his hands and there isn't anybody in his class who has more mechanical aptitude. He is also good at art. His best subject at school is math. Joe likes all of these things.

What should Joe do? Should he:

___ 1) look for an occupation in which he can use as many of his interests and abilities as possible?

___ 2) pick an occupation which uses math since there is a better future in that than in art or in working with his hands?

___ 3) decide which of these activities he is best at, or likes the most, and then pick an occupation which uses that kind of activity?

___ 4) put off deciding about his future and wait until he loses interest in some of these activities?

90. Betty gets very good science grades but this isn't her favorite subject. The subject she likes best is art even though her grades in it are only average. Betty is most likely to do well in her future occupation if she:

___ 1) forgets about her interest in art since she is so much better in science.

___ 2) doesn't worry about the fact that she isn't very good at art, because if you like something you can become good at it.

___ 3) looks for an occupation which uses both art and science, but more science than art.

___ 4) looks for an occupation which involves both science and art, but more art than science.

91. Bob says he really doesn't care what kind of work he gets into once he leaves school as long as it is working with people. If this is all Bob cares about he is likely to make a bad choice because:

___ 1) this kind of work usually requires a college degree.

___ 2) employers usually hire girls for such work.

___ 3) people look down on men who work with people because such work is usually done by girls.

___ 4) occupations in which one works with people can be very different from each other in the abilities and interests which are needed.

COOPERSMITH

SELF-ESTEEM INVENTORY

RMC Research Corp.
2570 W. El Camino Real
Mountain View, CA 94040
415/941-9550

PRACTICE ITEMS

A. I like to watch TV. LIKE ME _____ NOT LIKE ME _____

B. I'm a good worker. LIKE ME _____ NOT LIKE ME _____

---

1. I spend a lot of time daydreaming. LIKE ME _____ NOT LIKE ME _____ (33)
2. I'm pretty sure of myself. LIKE ME _____ NOT LIKE ME _____ (34)
3. I often wish I were someone else. LIKE ME _____ NOT LIKE ME _____ (35)
4. I'm easy to like. LIKE ME _____ NOT LIKE ME _____ (36)
5. My parents and I have a lot of fun together. LIKE ME _____ NOT LIKE ME _____ (37)
6. I never worry about anything. LIKE ME _____ NOT LIKE ME _____ (38)
7. I find it very hard to talk in front of the class. LIKE ME _____ NOT LIKE ME _____ (39)
8. I wish I were younger. LIKE ME _____ NOT LIKE M _____ (40)
9. There are lots of things about myself I'd change if I could. LIKE ME _____ NOT LIKE ME _____ (41)
10. I can make up my mind without too much trouble. LIKE ME _____ NOT LIKE ME _____ (42)
11. I'm a lot of fun to be with. LIKE ME _____ NOT LIKE ME _____ (43)
12. I get upset easily at home. LIKE ME _____ NOT LIKE ME _____ (44)
13. I always do the right thing. LIKE ME _____ NOT LIKE ME _____ (45)
14. I'm proud of my school work. LIKE ME _____ NOT LIKE ME _____ (46)
15. Someone always has to tell me what to do. LIKE ME _____ NOT LIKE ME _____ (47)
16. It takes me a long time to get used to anything new. LIKE ME _____ NOT LIKE ME _____ (48)

17. I'm often sorry for the things I do. LIKE ME _____ NOT LIKE ME _____ (49)

18. I'm popular with kids my own age. LIKE ME _____ NOT LIKE ME _____ (50)

19. My parents usually consider my feelings. LIKE ME _____ NOT LIKE ME _____ (51)

20. I'm never unhappy. LIKE ME _____ NOT LIKE ME _____ (52)

21. I'm doing the best work that I can. LIKE ME _____ NOT LIKE ME _____ (53)

22. I give in very easily. LIKE ME _____ NOT LIKE ME _____ (54)

23. I can usually take care of myself. LIKE ME _____ NOT LIKE ME _____ (55)

24. I'm pretty happy. LIKE ME _____ NOT LIKE ME _____ (56)

25. I would rather play with children younger than me. LIKE ME _____ NOT LIKE ME _____ (57)

26. My parents expect too much of me. LIKE ME _____ NOT LIKE ME _____ (58)

27. I like everyone I know. LIKE ME _____ NOT LIKE ME _____ (59)

28. I like to be called on in class. LIKE ME _____ NOT LIKE ME _____ (60)

29. I understand myself. LIKE ME _____ NOT LIKE ME _____ (61)

30. It's pretty tough to be me. LIKE ME _____ NOT LIKE ME _____ (62)

31. Things are all mixed up in my life. LIKE ME _____ NOT LIKE ME _____ (63)

32. Kids usually follow my ideas. LIKE ME _____ NOT LIKE ME _____ (64)

33. No one pays much attention to me at home. LIKE ME _____ NOT LIKE ME _____ (65)

34. I never get scolded. LIKE ME _____ NOT LIKE ME _____ (66)

35. I'm not doing as well in school as I'd like to. LIKE ME _____ NOT LIKE ME _____ (67)

36. I can make up my mind and stick to it. LIKE ME _____ NOT LIKE ME _____ (68)

37. I really don't like being a boy -- girl. LIKE ME _____ NOT LIKE ME _____ (69)

38. I have a low opinion of myself. LIKE ME _____ NOT LIKE ME _____ (33)

39. I don't like to be with other people. LIKE ME _____ NOT LIKE ME _____ (34)

40. There are many times when I'd like to leave home. LIKE ME _____ NOT LIKE ME _____ (35)

41. I'm never shy. LIKE ME _____ NOT LIKE ME _____ (36)

42. I often feel upset in school. LIKE ME _____ NOT LIKE ME _____ (37)

43. I often feel ashamed of myself. LIKE ME _____ NOT LIKE ME _____ (38)

44. I'm not as nice looking as most people. LIKE ME _____ NOT LIKE ME _____ (39)

45. If I have something to say, I usually say it. LIKE ME _____ NOT LIKE ME _____ (40)

46. Kids pick on me very often. LIKE ME _____ NOT LIKE ME _____ (41)

47. My parents understand me. LIKE ME _____ NOT LIKE ME _____ (42)

48. I always tell the truth. LIKE ME _____ NOT LIKE ME _____ (43)

49. My teacher makes me feel that I'm not good enough. LIKE ME _____ NOT LIKE ME _____ (44)

50. I don't care what happens to me. LIKE ME _____ NOT LIKE ME _____ (45)

51. I'm a failure. LIKE ME _____ NOT LIKE ME _____ (46)

52. I get upset easily when I'm scolded. LIKE ME _____ NOT LIKE ME _____ (47)

53. Most people are better liked than I am. LIKE ME _____ NOT LIKE ME _____ (48)

54. I usually feel as if my parents are pushing me. LIKE ME _____ NOT LIKE ME _____ (49)

55. I always know what to say to people. LIKE ME _____ NOT LIKE ME _____ (50)

56. I often get discouraged in school. LIKE ME _____ NOT LIKE ME _____ (51)

57. Things usually don't bother me. LIKE ME _____ NOT LIKE ME _____ (52)

58. I can't be depended on. LIKE ME _____ NOT LIKE ME _____ (53)

This booklet was prepared by RMC Research Corporation, Mountain View, California for use under National Institute of Education Contract No. NIE-400-78-0021

Study of the CAREER INTERN PROGRAM

INTERNAL-EXTERNAL SCALE

RMC Research Corp.
2570 W. El Camino Real
Mountain View, CA 94040
415/941-9550

INTERNAL-EXTERNAL SCALE

NAME _____ DATE _____

DIRECTIONS:

The purpose of this short task is to determine how you feel about certain things.

Read each of the following paired statements. Which of the two statements do you agree with more? Circle that letter. Choose only one. (However, be sure to choose one of the paired statements for each item).

Example: 1.a. Most children should be punished by their mothers.
b. A child knows when he does something wrong.

1.a. Children get into trouble because their parents punish them too much.
b. The trouble with most children nowadays is that their parents are too easy with them.

2.a. Many of the unhappy things in people's lives are partly due to bad luck.
b. People's misfortunes result from the mistakes they make.

3.a. One of the major reasons why we have wars is because people don't take enough interest in politics.
b. There will always be wars, no matter how hard people try to prevent them.

4.a. In the long run people get the respect they deserve in this world.
b. Unfortunately, an individual's worth often passes unrecognized no matter how hard he tries.

5.a. The idea that teachers are unfair to students is nonsense.
b. Most students don't realize the extent to which their grades are influenced by accidental happenings.

6.a. Without the right breaks one cannot be an effective leader.
b. Capable people who fail to become leaders have not taken advantage of their opportunities.

7.a. No matter how hard you try some people just don't like you.
b. People who can't get others to like them don't understand how to get along with others.

8.a. Heredity plays the major role in determining one's personality.
b. It is one's experiences in life which determine what they're like.

9.a. I have often found that what is going to happen will happen.
b. Trusting to fate has never turned out as well for me as making a decision to take a definite course of action.

10.a. In the case of the well prepared student there is rarely if ever such a thing as an unfair test.
b. Many times exam questions tend to be so unrelated to course work that studying is really useless.

11.a. Becoming a success is a matter of hard work, luck has little or nothing to do with it.
b. Getting a good job depends mainly on being in the right place at the right time.

12.a. The average citizen can have an influence in government decisions.
b. This world is run by the few people in power, and there is not much the little guy can do about it.

13.a. When I make plans, I am almost certain that I can make them work.
b. It is not always wise to plan too far ahead because many things turn out to be a matter of good or bad fortune anyhow.

14.a. There are certain people who are just no good.
b. There is some good in everybody.

15.a. In my case getting what I want has little or nothing to do with luck.
b. Many times we might just as well decide what to do by flipping a coin.

16.a. Who gets to be the boss often depends on who was lucky enough to be in the right place first.
b. Getting people to do the right thing depends upon ability, luck has little or nothing to do with it.

17.a. As far as world affairs are concerned, most of us are the victims of forces we can neither understand nor control.
b. By taking an active part in political and social affairs the people can control world events.

18.a. Most people don't realize the extent to which their lives are controlled by accidental happenings.
b. There really is no such thing as "luck."

19.a. One should always be willing to admit mistakes.
b. It is usually best to cover up one's mistakes.

20.a. It is hard to know whether or not a person really likes you.
b. How many friends you have depends on how nice a person you are.

21.a. In the long run the bad things that happen to us are balanced by the good ones.
b. Most misfortunes are the result of lack of ability, ignorance, laziness, or all three.

22.a. With enough effort we can wipe out political corruption.
b. It is difficult for people to have much control over the things politicians do in office.

23.a. Sometimes I can't understand how teachers arrive at the grades they give.
b. There is a direct connection between how hard I study and the grades I get.

24.a. A good leader expects people to decide for themselves what they should do.
b. A good leader makes it clear to everybody what their jobs are.

25.a. Many times I feel that I have little influence over the things that happen to me.
b. It is impossible for me to believe that chance or luck plays an important role in my life.

26.a. People are lonely because they don't try to be friendly.
b. There's not much use in trying too hard to please people, if they like you, they like you.

27.a. There is too much emphasis on athletics in high school.
b. Team sports are an excellent way to build character.

28.a. What happens to me is my own doing.
b. Sometimes I feel that I don't have enough control over the direction my life is taking.

29.a. Most of the time I can't understand why politicians behave the way they do.
b. In the long run the people are responsible for bad government on a national as well as on a local level.

APPENDIX D

The Correction for Guessing: Valid and Invalid Applications

The purpose of this appendix is to attempt to clarify issues concerning application of the so-called correction for guessing--particularly as that correction was employed in the Gibboney Associates (1977) evaluation of the Career Intern Program.

Tinkelman (1971) provides an excellent discussion of the correction for guessing. As he points out, if a test taker responds in a purely random fashion to k test items each of which has n choices, the expectation is that he or she will answer k/n items correctly and $k - k/n$ items incorrectly. If one assumes that all of the items answered incorrectly, W, were items which the respondent answered randomly, then $W = k - k/n$. It follows that $k/n = W/(n - 1)$. Since the total number of items answered correctly, R, is made up of the items to which the respondent knew the answer plus those which he or she got right by random guessing (k/n), the number of items to which the respondent knew the answer is given by R minus the correction for guessing $W/(n - 1)$.

What is important to note about the correction for guessing is that it is mathematically correct only when respondents answer correctly all items to which they know the answers and perform in a random fashion on all other items they attempt. As Tinkelman correctly points out, when guessing is non-random, the formula breaks down. It does not "work," for example, if the respondent is able to eliminate one or two of the answer choices as definitely incorrect and guesses among the remaining choices, or if he or she falls into a trap rigged by the ingenious item writer. It also does not work, as will be illustrated below, if the respondent guesses randomly on items where he or she knows the answer.

In the Gibboney study, the correction for guessing was applied to the "raw" test scores because "many of the people in the control group were completing the items by pattern responses on the answer sheet rather than by solving the problems and choosing their answer from among the distractors" (Vol. II, p. 16). As additional evidence that random responding occurred, the report indicated that (a) the increase in number of reading test items attempted from pre- to posttest was greater for the control than for the CIP group; (b) although control-group members attempted an average of 13.7 more items on the reading posttest than on the pretest, the number of items answered correctly increased by only .5; and (c) on the math test, the percentage of attempted items answered correctly increased from pre- to posttest for the CIP group but decreased for the controls.

These facts all suggest that members of the control group did, in fact, exhibit more random behavior (guessing) than members of the treatment group in responding to the posttest instruments. The critical question, as will be seen later, is whether they guessed only on items for which they could not have worked out the

correct answers or whether they also guessed on items they could have answered correctly. Under the former condition, the correction for guessing will serve its intended function whereas, under the latter condition, it will not. In fact, where guessing has occurred on items that could have been answered correctly, the correction for guessing will distort rather than correct. It will spuriously inflate differences between the guessing and the non-guessing groups.

Consider Ms. Ceebar, who knows the answer to 12 items on a 40-item test but has no idea what the correct answer may be to any of the remaining 28 items. If she responds only to those items about which she is knowledgeable, her score, 12, correctly informs us of the number of items to which she knows the answer. If she had answered the items correctly about which she was knowledgeable and had guessed on the rest, we would expect her to have answered 12 + 28/4, or 19 items correctly. Without a correction for guessing, we might mistakenly have assumed that she knew the answers to 19 items. If we apply the correction for guessing, however, we learn that, even though she answered 19 items correctly, she only knew the correct answers to 12 of them (19 - 21/3 = 12).

Now suppose that Ms. Ceebar was in a hurry and knew that she had nothing to gain from putting forth her best effort on the test. Rather than taking time to read and think about the items, she decided to save time and effort by simply marking her answer sheet at random. Under these circumstances she would (if she were average) have answered 10 items correctly and 30 items incorrectly. We might mistakenly have assumed, from this information, that she knew the answer to 10 items (actually she knew the answers to 12 items). If we apply the correction for guessing under these circumstances, Ms. Ceebar's corrected score is 10 - 30/3 or 0. Accepting this "corrected" score as a true indication of the number of items to which she knew the answers would lead us to a far more erroneous impression of her achievement level than acceptance of her uncorrected (but still definitely incorrect) score.

Assume that Ms. Ceebar was the average member of the control group and that she responded to the posttest in the manner just described. Mr. Teabar, who was the average member of the treatment group, also knew the answers to 12 items on the test. He answered these items correctly and guessed on the remaining 28 items. His score was 12 + 28/4 or 19. Ms. Ceebar's score was ten. Since the treatment effect is measured by subtracting the posttest score of the control group's average member from the posttest score of the treatment group's average member, we would conclude (erroneously) that the treatment had an impact of nine units (19 - 10 = 9).

Suppose we now correct both scores for guessing. Ms. Ceebar's score becomes 0 and Mr. Teabar's score becomes 12. Since 12 - 0 = 12, we now conclude that the effect of the treatment was a 12-point gain. This gain estimate is 33% larger than the gain estimate derived from scores that were not corrected for guessing. Both estimates are infinitely larger than the true gain that is obtained by subtracting the number of items to which Ms. Ceebar knew the answers (12) from the number of items to which Mr. Teabar knew the answers (also 12), 12 - 12 = 0.

The mathematics of the preceding argument are clear, but the argument itself may not apply exactly to the Gibboney Associates evaluation. Nevertheless, if there was even one more guess in the control group on an item the respondent could have answered correctly (by applying more time or effort) than there was in the treatment group, some distortion was introduced by the "correction" for guessing.

As pointed out earlier, the Gibboney Report provides ample and very convincing evidence that there was more random responding on the posttest among control group members than among treatment group members. The data show, in fact, that control group members, who correctly answered 71% of the items they attempted on the pretest, answered only 55% correctly on the posttest. The corresponding figures for the treatment group were 68% and 66%, respectively.

The Gibboney data also strongly suggest that some of the random responding occurred on items that the respondents could have answered correctly if they had made the effort. This inference is based on the fact that the control group members responded to 13.7 more items on the reading posttest than they did on the pretest. By chance alone they should have gotten 3.4 of these items correct. Their posttest scores, however, increased by only .5 points over their pretest scores, indicating that they must have answered 2.9 items incorrectly on the posttest that they had answered correctly on the pretest. It seems most unlikely that this phenomenon could be the result of a real loss of reading ability, considering the age of the students and the length of the pre-to-posttest interval. Thus, while the <u>possible</u> existence of a real loss of reading ability must be acknowledged, the probability that control group members responded randomly to some items that they could, with more effort, have answered correctly seems overwhelmingly greater.

The situation appears to be almost identical to the hypothetical example presented above involving Ms. Ceebar and Mr. Teabar. Random responding in the control group produced an uncorrected (for guessing) estimate of gain that was spuriously

high. Applying the correction for guessing, rather than correcting this problem, actually exacerbated it by making the already too-large estimate even larger.

The authors feel that the preceding discussion has made a convincing case against correcting scores for guessing under circumstances such as were observed in the Gibboney evaluation.

REFERENCES

Campbell, D. T., & Boruch, R. F. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), Evaluation and experiments: Some critical issues in assessing social programs. New York: Academic Press, 1975.

Campbell, D. T., & Erlebacher, A. E. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), Disadvantaged child. Vol. 3. Compensatory education: A national debate. New York: Brunner/Mazel, 1970.

Cook, T. D. & Campbell, D. T. Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally, 1979.

Coopersmith, S. The antecedents of self-esteem. San Francisco: W. H. Freeman & Co., 1967.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., Walker, D. F., & Weiner, S. S. Toward reform of program evaluation. San Francisco, CA: Jossey-Bass Inc., 1980.

Fetterman, D. M. Study of the Career Intern Program. Final technical report--Task C: Program Dynamics: Structure, Function, and Interrelationships. Mountain View, CA: RMC Research Corporation, 1981. [UR 465]

Gibboney Associates, Inc. The Career Intern Program: Final report, Volumes I and II. Blue Bell PA: Author, 1977. (NIE Papers in Education and Work)

Kenny, D. A. A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. Psychological Bulletin, 1975, 82(3), 345-362.

National Institute of Education. Compensatory education study, Final Report. Washington DC: Author, 1978.

National Institute of Education. Request for proposal: Study of the Career Intern Program. Washington, D.C.: Author, 1978. [RFP NIE-R-78-0004]

Porter, A. C. The effects of using fallible variables in the analysis of covariance. Unpublished doctoral dissertation. University of Wisconsin, 1967.

Raven, J. C. Matrix tests. Mental Health, 1940, I, 10-18.

Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs, 1966, 80, No. 609

Saretsky, G. The OEO P. C. experiment and the John Henry effect. Phi Delta Kappan, 1972, 53, 579-581.

Super, D. E. Career development. In J. R. Davitz & S. Ball (Eds.), Psychology of the educational process. New York: McGraw-Hill, 1970.

Tallmadge, G. K. Cautions to evaluators. In M. J. Wargo & D. R. Green (Eds.), Achievement testing of disadvantaged and minority students for educational program evaluation. Based on the proceedings of a U.S. Office of Education Invitational Conference. CTB/McGraw-Hill, 1978, 331-384.

Tallmadge, G. K. An empirical assessment of norm-referenced evaluation methodology. Mountain View, CA: RMC Research Corporation, 1981.

Tallmadge, G. K., & Horst, D. P. A procedural guide for validating achievement gains in educational projects. Washington, D.C.: U. S. Government Printing Office, 1976. (Stock No. 017-080-01516)

Tallmadge, G. K., & Wood, C. T. User's Guide: ESEA Title I evaluation and reporting system. Mountain View CA: RMC Research Corporation, 1976.

Tallmadge, G. K., & Yuen, S. D. Study of the Career Intern Program. Interim technical report No. 2--Task B: Assessment of intern outcomes. Mountain View, CA: RMC Research Corporation, 1980. [UR 462]

Thomas, T. C., & Pelavin, S. H. Patterns in ESEA Title I reading achievement. Menlo Park CA: SRI International, 1976.

Tinkleman, S. N. Test design, construction, administration, and processing. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Treadway, P. G., Stromquist, N. P., Fetterman, D. M., Foat, C. M., & Tallmadge, G. K. Study of the Career Intern Program. Final report Task A: Implementation. Mountain View, CA: RMC Research Corporation, 1981. [UR 464]

Winer, B. J. Statistical principles in experimental design. (2nd ed.) New York: McGraw-Hill, 1971.